Does syntax bias serial order reconstruction of verbal short-term memory?

Timothy Jones

National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care West (NIHR CLAHRC West) at University Hospitals Bristol NHS Foundation Trust; and University of Bristol

Simon Farrell

University of Western Australia

Author Note

Timothy Jones: National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care West (NIHR CLAHRC West) at University Hospitals Bristol NHS Foundation Trust, UK and School of Experimental Psychology, University of Bristol; Simon Farrell: School of Psychological Science, University of Western Australia.

This research was supported by Economic and Social Research Council (ESRC) grant RES-062-23-1199 whilst the first author was a PhD student at the University of Bristol, and the second author was a Reader at the University of Bristol. The first author's time was supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care West (NIHR CLAHRC West). The views expressed in this article are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

Correspondence should be addressed to Simon Farrell, School of Psychological Science, University of Western Australia, 35 Stirling Hwy, Crawley WA 6009. Email: simon.farrell@uwa.edu.au. Web: http://psy-farrell.github.io.

Abstract

Existing models of short-term sequence memory can account for effects of long-term knowledge on the recall of individual items, but have rarely addressed the effects of long-term sequential constraints on recall. We examine syntactic constraints on the ordering of words in verbal short-term memory in four experiments. People were found to have better memory for sequences that more strongly conform to English syntax, and that errors in recall tended to make output sequences more syntactic (i.e., a syntactic bias). Model simulations suggest that the syntactic biasing in verbal short-term recall was more likely to be accounted for by a redintegration mechanism acting over multiple items in the sequence. The data were less well predicted by a model in which syntactic constraints operate via the chunking of sequences at encoding. The results highlight that models of short-term memory should be extended to include syntactic constraints from long-term representations—most likely via redintegration mechanisms acting over multiple items—but we also note the challenge of incorporating such constraints into most existing models.

Keywords: working memory; chunking; redintegration; long-term knowledge; syntax

Does syntax bias serial order reconstruction of verbal short-term memory?

Introduction

Short-term memory refers to our ability to temporarily store a small amount of information in order so that it may be later re-used or processed further. Although often conceived as a separate system or buffer, it is clear that verbal short-term memory is not entirely independent from long-term language representations (Baddeley, 2000; Cowan, 1999). At the level of individual items, empirical findings show that words are remembered better than non-words or unknown words (e.g., Brener, 1940; Gathercole, Frankish, Pickering, & Peaker, 1999; Hulme, Maughan, & Brown, 1991; Hulme, Roodenrys, Brown, & Mercer, 1995; Patterson, Graham, & Hodges, 1994; Saint-Aubin & Poirier, 2000); memory is better for frequently occurring words than infrequently occurring words (e.g., Gregg, Freedman, & Smith, 1989; Hulme et al., 1997; Poirier & Saint-Aubin, 1996; Roodenrys, Hulme, Alban, Ellis, & Brown, 1994; Watkins, 1977), and for frequently occurring letters compared to infrequently occurring letters (e.g., Mayzner & Shoenberg, 1964); and that highly imageable and concrete words are more memorable than those that are more abstract (e.g., Bourassa & Besner, 1994; Walker & Hulme, 1999). When considering groups of items, pairings of letters that frequently co-occur in the English language are better remembered than pairings of letters that don't frequently co-occur (e.g., Baddeley, 1964), even when controlling for individual letter frequency (e.g., Kantowitz, Ornstein, & Schwartz, 1972; Mayzner & Shoenberg, 1964).

Our focus here is on the contribution of syntactic constraints to order memory. Previous work has established that short-term memory is sensitive to the relationships between words in a sequence. Sequences of words that form syntactic patterns are remembered better than re-orderings of the same words that do not form syntactic patterns (e.g., Epstein, 1961; Marks & Miller, 1964), and word pairs are remembered more accurately when the order of each pair conforms to syntactic rules (e.g., itchy window) than when the order of each pair is reversed so that it doesn't match syntactic rules (e.g., window itchy; Perham, Marsh, & Jones, 2009). Long-term language

representations therefore place *syntactic constraints* on the ordering of items in verbal short-term memory. Whilst this is a robust finding in the literature (e.g., Epstein, 1961; Marks & Miller, 1964; Perham et al., 2009), there is still no detailed explanation how sequence-wide relationships affect the ordering of individual items. Here we examine and test possible explanations for syntactic constraints on the ordering of words in verbal short-term memory. In particular, we show that syntactic constraints are best understood as a bias towards more syntactic word orderings, and explore how well this bias can be explained by two potential mechanisms: redintegration and chunking.

How do syntactic constraints have an effect on verbal short-term memory?

For syntactic constraints to influence verbal short-term memory, there must be some mechanism or representation that acts over multiple items, such that recall of items is dependent on the relationship of those items to others in a sequence. This in itself presents a challenge to contemporary models of ordering in short-term memory. Many existing models represent the order of items using positional markers (e.g., Brown, Neath, & Chater, 2007; Brown, Preece, & Hulme, 2000; Burgess & Hitch, 1999, 2006; Farrell, 2012; Henson, 1998; Lewandowsky & Farrell, 2008): each item is associated with a position marker, and items are retrieved one at a time, by successively cueing for each item with its associated positional marker. Positional marking is necessary to explain grouping errors (e.g., Henson, 1999) and protrusion errors (Henson, 1998; Henson, Norris, Page, & Baddeley, 1996), whereby items mistakenly recalled in the wrong group or trial tend to retain their within-group or within-trial position. However, the serial (one-at-a-time) retrieval associated with position marking does not obviously allow for linguistic effects that act over multiple items.

To identify possible routes for the incorporation of sequence-wide constraints into popular models of serial recall, we consider here two principle mechanisms by which sequential constraints might play a role in serial recall models: *chunking* and *redintegration*. This is not to suggest that no other mechanisms are involved, but consideration of how models account for other short-term memory

phenomena—particularly those related to effects of long-term memory—suggests that chunking and redintegration are the most likely candidates in providing sequential constraints on short-term recall over a broad range of contexts.

Chunking. Chunking implies the recruitment of robust, long-term representations for familiar groupings of items or events (e.g., Baddeley, 2000; Cowan, 1999; Miller, 1956). Chunks are unitized representations, with a high degree of association within chunks, and weak associations between chunks. Alternatively, chunking can be conceptualised as the compression of information on the basis of known codes (Mathy & Feldman, 2012). If short-term memory is limited to hold a certain number of chunks of information (e.g., Miller, 1956), forming several items into a single chunk means that more items can be stored in total. Although no formal description of this process has been presented, several qualitative descriptions have been offered in previous theories. Baddeley (2000) describes an episodic buffer, where information from separate short-term and long-term stores can be combined to form a single event or chunk. Cowan (1999) describes short-term memory as a highly activated portion of long-term memory, with a limited number of items activated at an above-baseline level at any one time. Items that are already strongly associated in long-term memory require less attention to co-activate than those that aren't strongly associated, leaving spare attentional resources to activate more items at an above-baseline level. A key finding addressing these theories as models of verbal chunking is that articulatory suppression does not modulate the sentence superiority effect (Baddeley, Hitch, & Allen, 2009), suggesting that the binding process involved in the formation of sentential chunks is not attentionally demanding.

The question here is how people specifically chunk on the basis of syntax so as to produce superior recall for more sentence-like sequences. One model of syntactic enhancement is that participants parse incoming word sequences according to the grammar of their native language (in this case, English). Under this model, sequences forming grammatical phrases are encoded as phrases rather than individual words. If grammar rules provide syntactic constraints on short-term memory, we might expect a

verbal sequence to be chunked according to how it fits with those rules. For example, the phrase: *eats red soup the brown mole*, should be chunked into a *verb phrase* (eats red soup) and a *noun phrase* (the brown mole), according to a simple hierarchical phrase structure (e.g., Chomsky, 1965; Pinker, 1998). It is harder to predict exactly how a verbal sequence might be chunked according to frequency-of-occurrence statistics. Cowan (2001) suggested that groupings of items within a chunk would be more strongly associated to each other (possibly due to frequency of co-occurrence) than groupings across chunk boundaries, but it is difficult to specify exactly what thresholds of association qualify items for inclusion in or exclusion from a chunk. For this reason, the current study only aims to test the hypothesis that verbal sequences are chunked according to grammar rules.

Previous work has noted the possibility that grammatical structures are chunked in memory. Gilchrist, Cowan, and Naveh-Benjamin (2008) found that recall of word sequences was limited by the number of clauses those sequences contained. In addition, they found that older adults showed a reduced tendency to access new clauses, but having accessed a clause were as likely to complete the clause as younger adults. Gilchrist et al. (2008)'s tentative interpretation was that words in the same clause form part of a single chunk in working memory, and thus present similar constraints as other types of chunks such as pre-learned word pairs (Naveh-Benjamin, Cowan, Kilb, & Chen, 2007). The dynamics of recall are also consistent with chunking of grammatical structures, with longer latencies to the first item in each constituent syntactic chunk (e.g., J. G. Martin, 1967; Wilkes & Kennedy, 1969).

There is a subtle distinction to be made here between chunking and grouping, although both can have similar effects on response latencies. Grouping involves the hierarchical organisation of items in a sequence according to perceptual qualities of the sequence at presentation. For example, sequences can be separated into groups by inserting temporal pauses between groups (e.g., McLean & Gregg, 1967; Parmentier & Maybery, 2008), presenting items in different voices or from different spatial locations (e.g., Parmentier & Maybery, 2008), or spontaneously by the participant (e.g., Farrell &

Lelièvre, 2009; Madigan, 1980). Response latencies to the first item in each group tend to be longer than for the other items in the group (e.g., Farrell & Lewandowsky, 2004; Maybery, Parmentier, & Jones, 2002; McLean & Gregg, 1967; Parmentier & Maybery, 2008). Chunking is a similar form of hierarchical organisation of items in a sequence, but based on unitized representations for each chunk (Johnson, 1970). For example, if a particular sub-sequence of items is very familiar, it could form a chunk in short-term memory (Baddeley, 2000).

Two testable predictions follow from a syntactic chunking mechanism. The first is that chunking according to the syntactic structure of the presented sequence at encoding should lead to improved recall for sequences that match the syntactic constraints in long-term memory, as they would form fewer chunks to be remembered. The second prediction is that latencies to the first item in each syntactic chunk should be longer than latencies to later items in each syntactic chunk. Previous work has found that people leave pauses between chunks in their recall (e.g., Ericcson, Chase, & Faloon, 1980), and chunking models of working memory assume a time cost to accessing new chunks that is borne out empirically (e.g., Anderson & Matessa, 1997; Daily, Lovett, & Reder, 2001; Johnson, 1972).

Redintegration. Redintegration is a process of reconstruction of degraded short-term memories using long-term knowledge (e.g., Brown & Hulme, 1995; Lewandowsky & Farrell, 2000; Schweickert, 1993). When the degraded short-term memory for a sequence is ambiguous (i.e., it could match several possible sequences), the reconstruction that is most likely to be recalled is the sequence (out of those credible options) that best matches the sequential constraints represented in long-term memory.

Redintegration has traditionally been applied to the reconstruction of individual items on a list to be remembered (e.g., Lewandowsky & Farrell, 2000; Schweickert, 1993). Nonetheless, Schweickert (1993) noted that redintegration may take place over a whole list of items, even though he only applied it to individual items. There are two areas of evidence that suggest redintegration does occur over multiple items in a list: The composition of the whole list seems to influence accuracy of recall of individual

items; and ordering within the list appears to be regularised.

The majority of evidence for whole-list composition influencing accuracy of recall of individual items comes from experiments comparing words and non-words (e.g., Hulme, Stuart, Brown, & Morin, 2003; Jefferies, Frankish, & Lambon Ralph, 2006; Patterson et al., 1994). In all cases, words are remembered better than non-words, but a telling finding is that non-words in lists mixed with words are recalled more accurately than non-words in pure non-word lists (e.g., Hulme et al., 2003; Jefferies et al., 2006). More pronounced effects are observed for frequency, where the recall of words is dependent on the frequency of other words, to the extent that high- and low-frequency words are recalled equally well on mixed lists (Hulme et al., 2003). Evidence from patients with semantic dementia shows that, when word meanings have been forgotten, there are many more phoneme migrations between the different unknown words (e.g., Patterson et al., 1994). Participants without such dementia demonstrate similar errors for lists of non-words (e.g., Jefferies et al., 2006). This has been taken as evidence that semantic representations help to glue together the phonemes within words (e.g., Jefferies et al., 2006; Patterson et al., 1994). Jefferies et al. (2006) also demonstrated that the recall of non-words in mixed lists was improved with an increasing number of words on the list, and improved with the frequency and imageability of the words. This suggests that not only do semantic characteristics influence the recall of words, but in making words more coherent, semantic constraints reduce phoneme migrations from non-words on the same list (Jefferies et al., 2006). With a mechanism of item-by-item redintegration, there is no particular reason to expect recall of non-words to be improved in mixed lists. Patterson et al. (1994) and Jefferies et al. (2006) concluded that there was a network of interactive activation (e.g., Dell & O'Seaghdha, 1992; N. Martin & Saffran, 1997) involving phonological and semantic representations. This network of activation is inherently noisy, and requires a process of 'cleaning up' to provide the final response, which must act over all of the items in the list (e.g., Jefferies et al., 2006).

A critical piece of evidence that redintegration may apply over multiple items at a

time, and one that is a central topic of this study, is regularisation in the ordering of items. In free recall with lists of six nouns, six adjectives, and six verbs (18 words altogether), Stanners (1969) showed that responses tended to include more grammatical groupings of words than would be expected by chance. For example, 'adjective-noun' pairings occurred at above chance levels (Stanners, 1969), demonstrating regularisation according to syntactic constraints in verbal short-term memory. In another study using immediate serial recall of sequences of six non-words, Botvinick and Bylsma (2005) found that participants trained on an artificial grammar for the non-words produced more regularisation errors with respect to that grammar than untrained (control) participants. Hoffman, Jefferies, Ehsan, Jones, and Lambon Ralph (2012) also showed that neurological patients struggling with semantic executive control still demonstrated a recall advantage for semantically related sequences and syntactically correct sequences, suggesting that intact executive control is not necessary to benefit from such constraints (e.g., Allen & Baddeley, 2009; Baddeley et al., 2009; Jefferies, Lambon Ralph, & Baddeley, 2004). The patients had a high tendency to regularise jumbled sequences into a valid syntax. Hoffman et al. (2012) concluded that linguistic constraints are automatically activated in these tasks, and one role of executive control in healthy participants is to inhibit these constraints in order to correctly recall lists of jumbled words. Together, these studies indicate that people have an automatic tendency to use their long-term knowledge of ordering of items to bolster recall through a redintegration process. If a redintegration mechanism is used in verbal short-term memory, we should expect to see some biasing or regularisation in recall. The pattern of regularisation taking place should then provide a clue to the source of sequential constraints held in long-term memory.

Our goal here is to provide a fine-grained examination of chunking and redintegration as determinants of the regularisation of word sequences. First, as part of Experiment 1, we describe an *iterated learning* technique used to magnify any cognitive biases in verbal short-term recall. We also discuss the metrics used to measure conformance to syntactic rules and frequency of occurrence of syntactic patterns in

everyday English, and how they are related. Following this, we present four verbal short-term memory experiments that manipulate the conformance to syntax of presented sequences and use an order reconstruction paradigm to explore the effects of syntax and frequency-of-occurrence on the accuracy and regularisation of the responses. Finally we present computational modelling of chunking and redintegration mechanisms within a constant framework of ordering in short-term memory (the Start-End Model; Henson, 1998), to explore how well these mechanisms fit the empirical findings.

Experiment 1

Experiment 1 employed an iterated learning technique to magnify any biases in verbal short-term recall, making them easier to detect and explore. Sir Frederic Bartlett (1932) pioneered the use of iterated learning in his 'serial reproduction' technique, used to investigate biases in memory for stories and pictures. In the pictorial version of this method, the first participant would be shown a drawing, and then asked to recreate it from memory after a 15-30 minute filled delay. A second participant would then be shown the first participant's new drawing, and asked to recreate this from memory after a filled delay, and so on. Bartlett explained the changes made to the drawing over a long chain of people as being due to a combination of imperfect memory and the use of commonly-held *schemas* in long-term memory to fill in the blanks.

Whilst Bartlett's (1932) serial reproduction experiments were informative, entertaining, and a landmark in memory research, he has been criticized for the lack of control in his experiments, the subjectiveness of his interpretations, and an absence of quantitative analyses (e.g., Mesoudi, 2007); indeed attempts at replicating his findings have failed (Carbon & Albrecht, 2012). However, in recent years, Bartlett's ideas have re-emerged and been applied with greater scientific rigour in the form of iterated learning. Griffiths and Kalish (2005, 2007) provided a formal justification, using Bayesian principles, for the idea that inherent cognitive biases will shape the languages being used as languages are passed from one generation to the next: an inter-generational version of iterated learning. Xu and Griffiths (2010) applied similar

mathematical techniques to iterated reconstruction memory, where one person's reproduction is passed to the next person to be remembered (e.g., Bartlett, 1932). They suggested that memories are reconstructed by combining a degraded memory trace with prior knowledge about stimuli, according to Bayesian inference (e.g., Hemmer & Steyvers, 2009; Huttenlocher, Hedges, & Vevea, 2000). After several iterations of reconstruction, reproduced stimuli should come to represent a sample from the prior probability distribution (representing inherent cognitive biases for those stimuli).

In an empirical demonstration of these Bayesian principles, Xu and Griffiths (2010) trained two groups of participants to recognize fish belonging to a particular category. The widths of the fish had the same variance in both groups, but a lower mean width in condition A than in condition B. The participants then completed many experimental trials of reconstruction of memory for the width of a schematic fish (e.g., Huttenlocher et al., 2000). The responses of one participant were presented to the next participant to study, over a chain of eighteen participants in each condition. Remembered fish widths converged towards a lower mean width in condition A than condition B over the chain of eighteen participants, as should be expected if iterated learning reveals the trained priors.

In previous memory reconstruction studies investigating biasing (Hemmer & Steyvers, 2009; Huttenlocher et al., 2000; Xu & Griffiths, 2010), the stimuli were very simple: individual objects that could vary only along a single dimension (e.g., widths of fish). Experiment 1 here applied the iterated learning technique to short-term memory for the ordering of sequences of seven words. The use of serial recall allowed an exploration of sequential biases (e.g., constraints on which items follow other items) in short-term memory. Word sequences were chosen so that different orderings of the same seven words could form sequences with varying conformance to syntactic rules, from a complete sentence to seven separate words. This allowed a more graded investigation of the relationship between conformance to syntax and accuracy than for pairings of words (Perham et al., 2009). The potential sentences were also semantically coherent (e.g., *little brown owl likes fat juicy mice*), in an attempt to facilitate any biasing by

background knowledge of syntax. The use of iterated learning should magnify any cognitive biases with regards to word ordering (syntax) in short-term memory, with sequences converging according to the cognitive constraints being imposed.

Participants

Twenty participants, 8 men and 12 women between the ages of 19 and 35 years, took part in Experiment 1. All were native English speakers and received a reimbursement of £5.

Materials

Eighty sequences of seven English words were generated by one of the authors such that each made a viable English sentence with the syntax: adjective adjective noun verb adjective adjective noun. Examples of these sentences are provided in the Appendix.

Metrics of conformance to syntax and previous experience

Conformance to syntax was assessed using four different measures, to ensure generality of findings. The first involved *parsing* a seven-word sequence using simple grammatical phrase structure rules (e.g., Chomsky, 1965; Pinker, 1998), whilst the other three—Dennis (*Dennis* 2009); Levenshtein (*Levenshtein* 1966); and Damerau (*Damerau-Levenshtein* 1964)—used different algorithms (from the domains of memorial distance, spell-checking and DNA comparison) to measure the minimum distance between a seven-word sequence and a whole sentence with appropriate syntax. As all four metrics produced similar results across experiments, and the parsing metric has some basis in linguistics, only the parsing metric is reported here. For the same reason, the parsing metric also forms the basis for much of the analysis and computational modelling in this study, although it could be replaced with other measures without significantly altering the findings.

The *parsing* measure involved parsing the word sequence into as few *parsed tokens* as possible. A parsed token is a combination of consecutive words that make a valid

syntactic group according to the simple grammar rules provided below (e.g., Chomsky, 1965; Pinker, 1998):

1. Noun-Phrase = Adjective* Noun

2. Verb-Phrase = Verb Noun-Phrase

3. Sentence = Noun-Phrase Verb-Phrase

   *Note: * = zero or more occurrences.*

For example, the sequence *[blind mole] [furry] [digs long tunnel] [thin]* contains four tokens, as shown by the square brackets. *[blind mole]* makes a noun phrase according to rule (1). *[long tunnel]* is a noun phrase by the same rule, but can be further combined into the verb phrase *[digs long tunnel]* due to rule (2). If the order of the words was *furry blind mole digs long thin tunnel*, rule (3) could also be used to make this only one parsed token (i.e., a complete sentence). It should be noted that fewer parsed tokens relate to greater conformance to English grammar.

To obtain a measure of people's previous experience of word sequences, the British National Corpus (British National Corpus Consortium, 2007) was analysed to gather frequency-of-occurrence statistics for the 105 unique permutations of seven tokens used in each trial of the current experiment:
adjective-adjective-noun-verb-adjective-adjective-noun. To estimate the influence of frequency-of-occurrence over the sequence as a whole (*whole-sequence* frequencies), a count was made of how often each permutation of the seven word-types appeared in sequence within the British National Corpus. In order to understand the influence of frequency-of-occurrence of particular pairings of word types, similar counts were made of how often each possible pairing of the tokens *adjective*, *noun*, and *verb* (9 possible pairings) appeared in order within the British National Corpus. The *pairing* frequency for each of the 105 possible permutations of seven words was the average of the constituent pairing frequencies for that permutation. For example, the pairing frequency for the permutation: *noun-adjective-adjective-adjective-verb-noun-adjective*;

would be the average of the frequency counts for the pairings: *noun-adjective*, *adjective-adjective*, *adjective-adjective*, *adjective-verb*, *verb-noun*, *noun-adjective*.

Design

Each trial consisted of seven English words taken from one of the sequences described in the *Materials* section above and provided in the Appendix. For the first participant, the order in which the words were presented was determined at random for each trial, with the constraint that there should be 16 sequences in each of 5 parsed tokens conditions: 2, 3, 4, 5 and 6 tokens; according to the parsing metric. The condition of '1 token' was excluded, as this would represent a viable sentence and it was felt that the presence of viable sentences might influence recall of other sequences by encouraging participants to search for sentential structure in all sequences. For balance, the condition of '7 tokens' (no syntactic relationship between the seven words) was also excluded.

For all participants following the first participant, the sequences presented at input were identical to the previous participant's response sequences (e.g., participant 5 would be presented with the exact response sequences of participant 4). The 80 trials were presented in a pseudo-random order in 4 blocks of 20 trials.

Procedure

Participants were tested individually. Each trial began with a fixation point (a cross) presented in the centre of the screen for 1000 ms, followed by a blank screen for 500 ms. The list items were then presented in black on a white background, one at a time in the centre of the screen for 500 ms each, with a 100 ms inter-stimulus interval when the screen was blank. The words were presented at this fairly rapid rate to limit opportunity for rehearsal, as rehearsal can compromise efforts to examine chunking (e.g., Chen & Cowan, 2009). Following presentation of the last list item, there was another blank screen for 500 ms, and then all 7 items were displayed on the screen in a random order in 3 rows centred in the middle of the screen: 3 items in the top row and 2 in each of the other rows. Participants were required to click on the items in the order

in which they were originally presented, using the left mouse button. When an item was clicked, it turned from black to grey, and it was not possible to click on it again. Once all 7 items had been clicked on, there was a blank screen for 1000 ms before the next trial began.

Four practice trials were presented before the experiment began to familiarise participants with the procedure. These were excluded from the data analysis. Participants were tested for a 45-minute session consisting of 4 blocks with 20 trials per block. Participants were encouraged to take a break after each block.

Data Analysis

There was an expectation of a monotonic increase in accuracy with conformance to syntax, and also some precedent that biasing should be stronger the less representative memorial stimuli are of cognitive biases (Hemmer & Steyvers, 2009; Huttenlocher et al., 2000), although it is not clear whether this applies to serial recall. For this reason, in tests of accuracy and biasing, the effect of parsing metric was coded as linear and quadratic contrasts. Contrasts provide a more appropriate and more powerful test of this monotonic relationship than an omnibus ANOVA. A significant linear contrast would indicate a linear relationship. Significant linear and quadratic contrasts would indicate a monotonic relationship that is not linear, or a non-monotonic relationship, depending on the coefficients of the linear and quadratic functions.

Results

Accuracy. The average accuracy on the task (items reported in their correct position) was 74% correct. A within-subjects one-way ANOVA demonstrated a significant effect of serial position on accuracy, Greenhouse-Geisser corrected $(\epsilon = .487) F(6, 114) = 53.324, p < .001, \eta_p^2 = .737$ (see Figure 1; left panel). Linear, $F(1, 19) = 80.107, p < .001, \eta_p^2 = .808$, and quadratic, $F(1, 19) = 45.016, p < .001, \eta_p^2 = .703$, contrasts were both significant, consistent with an extended primacy effect and small recency effect.

To examine the relationship between conformance to syntactic rules and frequency of occurrence of the presented sequences and recall accuracy for those sequences, Pearson's correlations were carried out for each participant. Table 1 (top section) gives the average correlations across the 20 participants, and also shows the results of one-sample t-tests assessing the significance of those correlations (with respect to a null correlation of 0). Greater conformance to syntax was related to higher frequency of occurrence, and both resulted in more accurate recall.

The relationship between the number of parsed tokens (i.e. syntactic chunks) in the presented list and average accuracy is shown in the right panel of Figure 1. The panel indicates that a greater number of tokens (i.e., less conformance to syntax) is related to less accurate recall. A within-subjects one-way ANOVA was conducted with number of parsed tokens (2 tokens to 6 tokens) as the independent variable, and proportion correct as the dependent variable. The linear contrast from the ANOVA was significant: $F(1, 19) = 7.28, p = .014, \eta_p^2 = .277$, but the quadratic contrast was not: $F(1, 19) = .181, p = .676, \eta_p^2 = .009$, and there were no other significant contrasts. This indicates a linear decline in accuracy as the number of parsed tokens in the presented sequences increased (i.e., as conformance to syntax decreased).

Latencies. A further analysis investigated response times for each item, to see whether people chunked sequences in memory into their constituent syntactic tokens (according to the parsing metric described above). The response sequence for each trial should best represent what was held in memory at the time of responding. However, for completeness, we also explored response times relating to syntactic patterns in the presented sequences (see Figure 2). As described in the introduction, if syntactic tokens are remembered as separate chunks, latencies to the first item in each syntactic token should be longer than latencies to later items in each syntactic token. For this reason, latencies for the first item and later items within each syntactic token in the sequences were compared. Items at the first and last serial positions were excluded, for the following reasons. Latencies for the first response in a sequence are usually much longer than for the other responses (e.g., Anderson & Matessa, 1997; Maybery et al., 2002). As

the first response is also always the first item in a syntactic token, if it were included the comparison could be biased towards longer times for first items in a syntactic token. The item at the last serial position should tend to have a faster latency than the others as it is the only item remaining, and this can only ever be either a singleton, or later than first in a syntactic token, so could also bias the comparison if included. Any syntactic tokens containing only a single item were also excluded.

After these considerations, four participants could not provide a complete set of latency data for the remaining serial positions for the presented or response sequences and were excluded from the analysis. Paired-samples t-tests showed that the average latency to the first position (1193 ms) and later positions (1130 ms) in each syntactic token in the presented sequences was not significantly different, $t(15) = 0.93, p = .367$, 95% CI = −81 ms; 207 ms, and neither was the average latency to the first position (1279 ms) and later positions (1153 ms) in each syntactic token in the response sequences, $t(15) = 1.97, p = .068$, 95% CI = −10 ms; 261 ms. The latency results are summarised further in the general discussion by aggregating the data from all experiments to provide a more powerful analysis.

Biasing. A key question is whether participants regularise sequences, and if so what pattern of biasing is produced. Figure 3 (left) shows how the sequences from each starting group (2 to 6 parsed tokens) change in their syntactic structure as they are passed along a chain of participants, according to the parsing metric. Participants are numbered on the x-axis according to their position in the chain of participants. The dashed horizontal line represents the expected number of parsed tokens for randomly jumbled sequences (i.e., as if errors were random). The sequences in Experiment 1 appear to converge as they are passed along the chain of participants. Whilst the lines on the graph (representing the mean number of parsed tokens in sequences in each of the different starting groups) do not end up directly on top of one another, the sequences have remained at roughly the same overall level of conformance to syntax for the second half of the experiment. Any remaining variation in conformance to syntax between the groups could be put down to variation amongst particular sequences of items rather

than a lack of convergence. After the final participant, the sequences contained fewer parsed tokens than might be expected if errors were random with respect to syntax (because the lines on the graph converge below the dashed horizontal line).

Figure 3 (right) demonstrates how the sequences change in frequency-of-occurrence as they are passed along the chain of participants. The whole-sequence and pairing frequencies use different scales (i.e., they have different maximum values), so were normalised by dividing by the maximum possible value in each case (the values still look different because the whole sequence frequencies are more positively skewed). The dashed lines are chance lines showing the expected mean values under the null hypothesis, when errors are unrelated to the syntax of the presented sequences (i.e. unbiased). As the measures are correlated, it is not surprising that the two lines look fairly equivalent, and finish at a frequency of occurrence greater than would be expected by chance.

Although Figure 3 is informative about the long-run effects of background knowledge, it is possible that just a few people make large regularizing changes to the sequences. More critically, the characteristics of the input sequences differed between participants. If errors were random with respect to syntax, we would expect highly syntactic sequences to become less syntactic, and less regular sequences to become more regular, causing a natural 'regression to the mean' in the number of parsed tokens. Accordingly, a method is required to test whether people generally changed the sequences with which they were actually presented in a way that differs from what would be expected by chance. To assess biasing in responding by individual participants, a bootstrap simulation was conducted.

The bootstrapping technique works as follows (see Figure 4):

1. We know the exact recall order for every trial for each participant. For example, the recall order on a particular trial might be 1-5-2-3-7-4-6, where the numbers refer to the input serial positions (the original presentation order) of the reported items (see Figure 4; left side).

2. We randomly mix up the actual recall orders from trials within each

parsed-tokens condition (the number of parsed tokens in the presented sequence for that trial); see Figure 4 (right side). The reason for mixing up recall orders within the same parsed tokens condition is so that the accuracy and serial position curve within that condition will remain identical to the real experiment for each participant.

3. We apply the mixed up, 'bootstrapped', recall orders to the actual presented word sequences, check the syntax of the resulting 'bootstrapped' responses, and calculate the change in parsed tokens between the presented sequences and the 'bootstrapped' responses.

4. We repeat from (2), in this case 1000 times, and average the relevant metrics (e.g. change in conformance to syntax) over the repetitions.

The bootstrap represents a simulation of how each participant would respond, given their actual working memory ability and recall error patterns, if their errors in recall were not linked to the syntax of the presented sequences. The bootstrapped sequences will be subject to the same regression to the mean as the sequences actually recalled, and so serve as a baseline for the observed data.

Figure 5 (left panel) shows the average change in number of parsed tokens between presented sequences and responses for groups of presented sequences with different numbers of parsed tokens (2 to 6), comparing the data to the bootstrap. It is clear that sequences which are low in conformance to syntax become more syntactic, and sequences which are high in conformance to syntax become less syntactic. This would be expected if errors were random with respect to syntax, as demonstrated by the bootstrap. The information is re-plotted in the right panel as the difference between the data and the bootstrap. A within-subjects one-way ANOVA was carried out with parsed tokens condition as the factor. The significant intercept indicated that the empirical responses had significantly fewer tokens (were more consistent with English syntax) than the bootstrapped responses, $F(1, 19) = 47.776, p < .001, \eta_p^2 = .715$. Although the linear contrast ($F(1, 19) = 3.829, p = .065, \eta_p^2 = .168$) approached significance (suggesting a trend towards stronger biasing for more syntactic sequences), both it and the quadratic contrast ($F(1, 19) = .379, p = .546, \eta_p^2 = .02$) were not significant.

Similar bootstrap simulations were run for the frequency metrics. A change score was calculated for both the data (where the change was between each input sequence and the corresponding response sequence) and the bootstrap (calculating the average change between the input sequence and the 1000 corresponding bootstrapped responses). Figure 6 shows the average change in conformance to syntax, and in whole-sequence and pairing frequency, between the presented sequences and responses for each participant in the real data and the bootstrap simulations. Paired-samples t-tests showed that, using each metric, the ordering of participants' responses were more common in English (according to the British National Corpus, British National Corpus Consortium, 2007) than expected by chance (see Table 2).

Discussion

Iterated learning was applied to order reconstruction of sequences of words that could potentially form a sentence. Replicating previous findings (Miller & Selfridge, 1950; Perham et al., 2009), word sequences were remembered more accurately when they had greater conformance to syntax. The format of the sequences used in the current study allowed for more variation in conformance to syntax than other studies (e.g., Perham et al., 2009). The results suggested a linear relationship between the number of parsed tokens in a sequence and memory accuracy.

When examining the syntax of sequences across generations, it was apparent that sequences did not converge towards full sentences as might be expected under a syntactic bias. It may be that participants are sensitive to the distributional properties of the stimuli presented to them during the experiment (e.g., Huttenlocher et al., 2000), such that the long-term knowledge constraining recall was a mixture of learning prior to the experiment, and learning in the experiment. As few of the sequences presented to participants in this experiment were full sentences, participants may have been reticent to produce full sentences as responses.

On first inspection, the presence of systematic biasing appears to provide evidence for a multiple-item redintegration mechanism, whereby degraded sequences are

reconstructed with reference to syntactic constraints in cognition. However, further reflection suggests that the appearance of a bias could be caused simply by more accurate memory for particular sequences, or particular sub-sections of sequences, through chunking at encoding. Better-remembered groupings of items (e.g., more syntactic groupings of words) are more likely to remain intact as they are passed along the chain of participants, whilst other groupings of items (e.g., less syntactic groupings of words) are more likely to be changed through errors. Errors might then be random (i.e., not systematically biased), but if they happen to improve the match between a sequence and a population's inherent cognitive constraints, the next participant will remember that sequence more accurately, and so it is less likely to undergo a transformation. This bears some similarity to a survival-of-the-fittest mechanism (Darwin, 1859/1985), with certain groupings of items more likely to survive a journey through the memory of each participant, due to being chunked together in memory. Such a mechanism could result in the appearance of a systematic bias towards background knowledge, even though errors in recall are not systematically biased. After presentation of several more controlled experiments, we address this possibility directly using computational modelling.

The latency data provided no evidence for chunking according to the parsed tokens in the presented sequences or the response sequences. If there is some mechanism of chunking according to syntactic rules, it appears that either it does not affect response latencies, or it does not follow syntactic rules as we have specified them. The presence of systematic biasing and the lack of any effect of syntactic structure on response latencies provides some argument against a chunking mechanism, but does not completely rule this explanation out. However, we defer any strong conclusions on this point until a later analysis, where the latency data from all four experiments are examined in aggregate.

Certain aspects of the iterated learning method limit the conclusions that we can draw from this experiment. Due to the nature of iterated learning, each participant was presented with different sequences with different characteristics, so the averages used in

the ANOVA exploring the effect of conformance to syntax on accuracy are based on differing numbers of observations for each participant. Also, the iterated learning paradigm uses the responses of one participant as stimuli for the next participant, which means that the data for each participant are not entirely independent of each other, breaking the parametric assumptions of t-tests on the correlations and the analysis of variance. The results are therefore indicative of a linear relationship between the number of parsed tokens and accuracy, but need to be replicated under more controlled conditions. Experiments 2, 3, and 4 serve this purpose, as well as testing for syntactic regularisation in verbal recall when the semantic relatedness between the words is varied.

Experiments 2, 3, and 4

The remaining Experiments were similar to Experiment 1, except that none used an iterated learning paradigm (i.e., the responses of participant n were not passed as stimuli to participant $n + 1$). In these experiments, the presented sequences were manipulated so that each participant received a particular number of sequences from each parsed-tokens condition. This allows a more controlled investigation of how the responses in each group differ from the presented sequences in terms of their conformance to syntax. In Experiments 2 and 3, there were an equal number of sequences (16) in each of 5 parsed-tokens conditions (2 to 6 tokens). Experiment 4 aimed to highlight the presence of syntax in the sequences by including more sequences with fewer parsed tokens.

In Experiment 1, the sequences had the potential to make sense semantically, and therefore a more syntactic ordering of the words might also induce more associations to semantic meanings. Semantics and syntax can have separable influences on the accuracy of recall (e.g., Marks & Miller, 1964), and it is possible that semantics also had an influence on biasing in verbal short-term recall in Experiment 1. Accordingly, a further change from Experiment 1 was the inclusion of sequences with less semantic relatedness.

Experiment 2 involved a manipulation of semantic relatedness between the words

within a sequence, with a *meaningful* group and a *meaningless* group of sequences. The sequences in the meaningless group could still conform to syntactic rules, but with little overall semantic cohesion between the constituent words, similar to the anomalous group of Marks and Miller (1964). In addition, the words in the 'meaningless' sequences had a reduced chance of having been experienced together before, meaning that any biasing effects could be more confidently credited to the grammatical relationships between items. Comparing results for the meaningful and meaningless groups allows an investigation of the impact of semantics and frequency of co-occurrence of specific words on accuracy and syntactic biasing in short-term recall.

Experiment 3 was almost identical to Experiment 2, but with all of the presented sequences belonging to the *meaningless* condition. If participants are sensitive to the characteristics of sequences presented during the experiment (e.g., Huttenlocher et al., 2000), there is a possibility that the presence of some meaningful sequences prompts or primes (Bock, 1986) people to re-order meaningless sequences in a similar fashion. The presence of meaningful sequences in the experiment might be a major determinant of the regularisation of the meaningless sequences. In order to assess the effects of syntax on the regularisation of word sequences with minimal influence from semantics, Experiment 3 replicated Experiment 2 using only meaningless sequences.

Experiment 4 was a repeat of Experiment 3, but with many more syntactic sequences presented, including full sentence structures. One possibility, suggested by an anonymous reviewer, is that people may not strongly respond to the syntactic nature of sequences if heterogeneous sequences are mixed together. In other words, the use of syntax may strategically be adjusted to the expected usefulness of syntax in representing sequences. The results of Baddeley et al. (2009), which suggest a more automatic use of language knowledge, speak against this possibility to some extent. Nonetheless, it may be that long-term knowledge is more implicitly recruited in response to the statistics of the experiment. To address this possibility, less regular sequences were intermixed with highly regular sequences (including sentences) in order to encourage the use of syntax.

Experiments 2, 3, and 4 were very similar in design, and for reasons of parsimony

are reported together.

Participants

Twenty participants, 8 men and 12 women between the ages of 19 and 49 took part in Experiment 2. Twenty participants, 3 men and 17 women between the ages of 19 and 29 took part in Experiment 3. Forty participants, 19 men and 21 women between the ages of 18 and 30 took part in Experiment 4. All participated either voluntarily or in return for course credit, and were native English speakers.

Materials, Design and Procedure

For Experiment 2, the design and procedure were identical to Experiment 1 except that the 80 sentences were randomly allocated to two groups of 40 for each participant: a *meaningful* and a *meaningless* group. The meaningful group was left exactly as it was. In the meaningless group, all of the nouns, verbs and adjectives were randomly mixed between sentences (within their own form class, so nouns mixed with nouns, etc.) This created 40 sequences from the same words, and with the same syntax: *adjective adjective noun verb adjective adjective noun*, but with the constraint that none of the 7 words was from the same original sentence. Examples of both types of sequences are presented in the appendix. For each participant, the order of the words in the presented sequences was randomly varied with the constraint that there were an equal number of sequences in each of five parsed-tokens conditions (2, 3, 4, 5, and 6 tokens), for both the meaningful and meaningless sequence types. The 80 trials were presented in pseudo-random order. Experiment 3 used the same procedure but with meaningless sequences. Experiment 4 used the same procedure but the make-up of parsed-tokens was as follows: 30 trials with 1 token; 20 trials with 2 tokens; 10 trials each with 4, 5, and 6 tokens.

Results

Accuracy. The mean accuracy on Experiment 2 was 63% correct on the meaningful sequences and 62% correct on the meaningless sequences; for Experiments 3

and 4 it was 65% and 63% respectively. Figure 7 (left panels) shows the serial position curves for Experiments 2 (top) to 4 (bottom). A within-subjects two-way ANOVA was conducted to investigate the effects of serial position and semantic group (meaningful/meaningless) on accuracy in Experiment 2; whilst a within-subjects one-way ANOVA was carried out to explore the impact of serial position on accuracy in Experiments 3 and 4. Results are reported in Table 3. There were significant linear and quadratic contrasts over serial position in all experiments, demonstrating an extended primacy effect and small recency effect. The serial position curves were similar for both semantic groups in Experiment 2 (i.e., there was no main effect of semantic group and no interaction).

The relationship between the number of parsed tokens and average accuracy in each condition is shown in the right panels of Figure 7. A within-subjects two-way ANOVA was conducted for Experiment 2 with number of parsed tokens (2 tokens to 6 tokens) as one factor, semantic condition (meaningful/meaningless) as the other factor, and proportion correct as the dependent variable. A one-way within-subjects ANOVA was carried out to investigate the number of parsed tokens in the presented sequences on accuracy for Experiments 3 and 4. Results are reported in Table 4, which shows that none of the effects were significant for Experiment 2. Experiment 3 found significant linear and cubic contrasts over the number of parsed tokens; there was a monotonic decline in accuracy with increasing numbers of parsed tokens, but this appeared to level off for the middle 3 conditions. Experiment 4 found significant linear and quadratic contrasts, showing a slight monotonic decline in accuracy which levelled off with more parsed tokens in sequences.

As for Experiment 1, Pearson's correlations were carried out between each of the measures and proportion correct for each experiment. Table 5 gives the average correlations, and also shows the results of one-sample t-tests assessing the significance of those correlations (with respect to a null correlation of 0). As expected, the parsing and frequency metrics were significantly correlated in all experiments. Experiment 2 found a significant average correlation between the frequency metrics and accuracy, whilst

Experiment 3 found a significant average correlation between the parsing metric and accuracy. All average correlations differed significantly from 0 in Experiment 4.

Latencies. Latency analyses were carried out separately for Experiments 2, 3, and 4 as for Experiment 1 (with semantic groups combined for Experiment 2). The top three rows of Figure 8 show the corresponding means as a function of serial position, and whether or not an item was the first in its chunk. Experiments 2 and 3 only produced significant effects of serial position, with no overall impact of first vs. later item in syntactic chunk, and no interaction between this and serial position. For Experiment 4, latencies were significantly affected by whether an item was at the initial (vs later) position in a chunk in the response sequences, and this interacted with serial position, with larger effects apparent at serial positions 3 and 4; only serial position was significant in the analysis of presented sequences. It is possible that the presence of more syntactic sequences in Experiment 4 could have encouraged syntactic chunking. Alternatively, it may be that latency effects were too small to be detected with the smaller sample sizes used in the earlier experiments. To address this issue, we ran an aggregated analysis of the latency data over the four reported experiments.

The aggregate analysis is summarised in Figure 8. Two-way within-subjects ANOVAs on the aggregated data examining the effects of serial position (2 to 6), and syntactic chunk position (first versus later) in the presented sequences and response sequences revealed significant effects of serial position, $F(4, 380) = 69.7, p < .001, \eta_p^2 = .423$ and $F(4, 380) = 81.3, p < .001, \eta_p^2 = .461$, respectively. For the presented sequences, the effect of syntactic position, $F(1, 95) = 3.61, p = .06, \eta_p^2 = .037$, and the interaction, Greenhouse-Geisser corrected ($\epsilon = .841$) $F(4, 380) = .96, p = .431, \eta_p^2 = .01$, were not significant. However, for the response sequences syntactic position, $F(1, 95) = 18.289, p < .001, \eta_p^2 = .161$, and its interaction with serial position, Greenhouse-Geisser corrected ($\epsilon = .897$) $F(4, 380) = 5.5, p < .001, \eta_p^2 = .055$, were significant. Figure 8 suggests that the effects involving syntactic position are driven by longer response latencies for the first items in a syntactic chunk only at the fourth (and possibly third) serial position, although this is

not consistent across experiments. One possibility raised by this analysis is that participants tend to spontaneously group (e.g., Farrell & Lelièvre, 2009; Madigan, 1980) the sequences in, say, a 3–4 or 4–3 grouping pattern , and that this grouping is amplified if syntactic chunks match their grouping pattern. It is noteable that these syntax-based latency effects were related to the syntax of response sequences; there was little evidence of grouping according to the syntactic chunks in the presented sequences.

Biasing. To examine the extent of biasing in people's responses, similar bootstrap simulations were run on the data as for Experiment 1; for Experiment 2 separate bootstraps were run for sequences in each semantic group. Figure 9 (left panels) shows the average change in number of parsed tokens between presented sequences and responses for the different types of presented sequences, in both the data and the bootstrap simulations. The rows in the figure correspond to Experiments 2–4 respectively. The negatively sloped lines in the left panels show that both the data and the bootstrap are sensitive to purely statistical constraints: reducing conformance to syntax in more syntactic sequences (since there is more opportunity for errors to make the sequence less syntactic), and increasing conformance for less syntactic sequences. The information is re-plotted in the right panel as the difference between the data and the bootstrap for each condition. A within-subjects two-way ANOVA was carried out with parsed tokens condition (2 to 6) and semantic group (meaningful/meaningless) as the factors for Experiment 2; whilst a within-subjects one-way ANOVA was conducted for Experiments 3 and 4 with parsed tokens condition as the only factor. Significant negative intercepts for all experiments indicated that the real responses had significantly fewer tokens than the bootstrapped responses: $F(1, 19) = 21.588, p < .001, \eta_p^2 = .532$, $F(1, 19) = 14.758, p = .001, \eta_p^2 = .437$, and $F(1, 39) = 68.58, p < .001, \eta_p^2 = .637$, for Experiments 2–4 respectively. The other effects are reported in Table 6. For Experiment 2, the significant interaction of the linear contrast over number of parsed tokens and semantic group demonstrated a steeper slope in the meaningless group. Analysis of the simple effects of parsed tokens condition for each semantic group showed a significant linear contrast for the meaningless group, $F(1, 19) = 7.186, p < .05$, but no

significant quadratic contrast, $F(1, 19) = 3.307, p > .05$. Neither contrast was significant for the meaningful group, $F(1, 19) = 1.815, p > .05$, and $F(1, 19) = .231, p > .05$.

Bootstrap simulations were also conducted for the metrics of frequency of occurrence. Figure 10 shows the average change in whole-sequence and pairing frequency, between the presented sequences and responses both in the data and the bootstrap simulation (results for the parsed tokens measure are also shown in the top row for comparison, but are redundant with the analysis just presented). Averages for the two semantic groups are presented for Experiment 2 (to make the figure easier to read), whilst averages for each participant are presented for Experiments 3 and 4. Table 7 describes the results of three separate 2×2 within-subjects ANOVA for Experiment 2, investigating the effects of semantic group (meaningful vs. meaningless) and data type (real vs. bootstrap) on the change between presented sequences and responses according to each metric. Table 8 gives the results of paired samples t-tests comparing the data to the bootstrap for each metric for Experiments 3 and 4. The real responses were more syntactic than would be expected by chance (i.e., compared to the bootstrap), and more frequent according to the British National Corpus (British National Corpus Consortium, 2007) than expected by chance, and this did not differ significantly between the semantic groups in Experiment 2 (i.e., there was no effect of semantic group and no interaction).

Discussion

Overall, more syntactic sequences (i.e., those with fewer parsed tokens) were remembered more accurately than less syntactic sequences, and there was a systematic bias towards sequences with greater conformance to syntax and a higher frequency of occurrence than would be expected by chance. However, Experiment 2 showed no relationship between conformance to syntax and accuracy. The lack of a relationship cannot be explained by floor effects, because even at the minimum of the serial position curves (see Figure 7; left panel), accuracy was still around 40%, which is well above chance levels; and the effect was found in Experiments 3 and 4 with similar overall

accuracy. While there might be some feature of Experiment 2 (i.e., the mixing of meaningful and meaningless sentences) that contributed to this discrepancy, this may well simply be a Type II error, especially as the effect observed in Experiments 3 and 4 is not numerically large. In drawing our conclusions below, we will be concerned with the overall pattern emerging across the experiments.

There was no impact of the manipulation of semantic meaning on accuracy. Previous studies found improved accuracy for whole sequences with semantic meaning over whole sequences with equivalent syntax but reduced semantics, using a similar method to reduce semantic meaning in the sequences (e.g., Marks & Miller, 1964; Miller & Isard, 1963). Also, Jefferies et al. (2004) found improved accuracy when several sentences made a coherent story than when the sentences were unrelated to each other. There are some possible explanations for the difference between the current results and these previous findings (Jefferies et al., 2004; Marks & Miller, 1964; Miller & Isard, 1963). Miller and Isard (1963) found their effect with a shadowing technique, whereby people had to concurrently dictate a long stream of words that they heard through headphones. They considered this more a perceptual effect than a memory effect. Marks and Miller (1964) demonstrated a semantic advantage both for sequences with a complete syntax and those with jumbled syntax (i.e., sequences of semantically-related jumbled words were remembered better than semantically-unrelated jumbled words). However, each trial entailed remembering 5 sequences of 5 words each, for a total of 25 words, substantially longer than the lists presented here. The Jefferies et al. (2004) study used a similarly high number of words: six sentences of five to eight words each. Some research suggests that verbal items are initially encoded phonologically, and gradually recoded semantically over time (Kintsch & Buschke, 1969; Tell, 1972). Memory for sets of larger numbers of words (e.g., Jefferies et al., 2004; Marks & Miller, 1964) may therefore incorporate different processes, more likely to involve semantic long-term memory, than memory for sequences of 7 words, and this could account for the different result in the current experiment. The relatively fast presentation rate used in the current experiment may also have limited the time available for semantic

processing (N. Martin & Saffran, 1997). Additionally, Jefferies et al. (2004), Marks and Miller (1964), and Miller and Isard (1963) all included tests of item memory, whereas our study did not. Potter and Lombardi (1990) demonstrated that semantics can enhance recall based on the overall gist of a sequence, and it may be that providing the items at recall limited the opportunity for observing effects of semantics. Another possibility is that the manipulation of semantic relatedness may not have been strong enough to produce an effect. However, the manipulation used in this experiment was very similar to that in Marks and Miller (1964) and Miller and Isard (1963), where an effect of semantics was found, arguing against this as a likely explanation. In any case, the results from Experiments 2, 3, and 4 show that the main focus of the experiments—the relation of long-term syntactic knowledge to performance and the nature of errors—was similarly observed irrespective of the semantic content of the sequences.

Generally, the findings support the conclusion that syntax affects short-term memory performance even when word sequences are not particularly meaningful, and groupings of specific words may not have been experienced together often in the past. In other words, at least some effect of syntax may occur because certain *types* of words have co-occurred together in the past, or match syntactic rules, rather than the specific words themselves (Smith, 2005). Chunking and redintegration mechanisms, as implemented in current computational models (e.g., Botvinick & Plaut, 2006; Burgess & Hitch, 2006), tend to be sensitive to known combinations of specific items (e.g., words); that is, they predict more accurate recall for particular orderings of particular words that have been experienced before.

Having established across several experiments that long-term sequential knowledge biases recall of sequences, we now turn to the question of if, and how, the chunking and redintegration mechanisms can account for the syntactic effects in our data. To answer this question, we implemented redintegration and chunking in a common serial ordering framework (Henson, 1998) and examined the extent to which the data generally conformed with the predictions of the two mechanisms.

Computational Modelling

The results of the experiments in this study point to two general effects of syntactic constraints on verbal short-term recall: a) serial reconstruction was more accurate for more regular sequences (though see Experiment 2); and b) recall errors were biased towards making a sequence more regular (as measured by the number of parsed tokens in the output sequence or its frequency of occurrence). At face value, these findings are consistent with the claim that long-term representations have a biasing effect on short-term memory through a redintegration mechanism, with improved accuracy of recall arising when memoranda match the constraints from long-term memory. However, as we explained in the discussion for Experiment 1, it is possible to entertain a version of the chunking hypothesis under which a biasing effect would emerge, especially where iterated learning was employed. Under such a model, if sub-sequences that form large chunks are less susceptible to perturbation, random perturbations on smaller chunks (or individual items) may well produce more regular sequences on average.

To ensure that chunking and redintegration mechanisms do indeed predict the behaviour attributed to them above Farrell and Lewandowsky (2010), the behaviour of both mechanisms was simulated within the same computational modelling framework, under a variety of parameter estimates. The modelling did not involve fitting the models directly to the data (e.g., using maximum likelihood). This was partly because it was not clear how best to combine the fits to the separate metrics of long-term knowledge effects (i.e., the effects on accuracy and change in number of parsed tokens), but also to account for any possible differences in flexibility between the models. Accordingly, the model analyses explore the the overall pattern of effects across a range of parameter values. For each set of parameter values in each model, predictions were derived from that model. By inspecting the distribution of predictions, we can see the extent to which a model systematically predicts an effect (i.e., how closely the predictions under different parameter values cluster together), and how well the data accord with the model's predictions on average. This means we can look at how well

the predictions correspond to the data on average, rather than focussing on how close a particular model can get under specific parameter values.

To keep the models as similar as possible (so any differences in behaviour could be uniquely attributed to the mechanism by which syntactic knowledge constrains verbal short-term memory), mechanisms were implemented in a representative model of serial recall, the Start-End Model (SEM; Henson, 1998). This model was chosen as it incorporates the majority of important assumptions on which serial recall models have converged to fit empirical findings (e.g., positional representations, primacy gradient Hurlstone, Hitch, & Baddeley, 2014), and is relatively simple and faster to run compared to alternative models (e.g., Brown et al., 2000; Burgess & Hitch, 1999, 2006; Lewandowsky & Farrell, 2008). We implement a simplified version of the model here that omits some specific mechanisms in Henson (1998) that are necessary to handle particular effects in the serial recall literature that are not of concern here.

The Start-End Model

Henson's (1998) SEM assumes that items are paired with a two-dimensional representation of position within the sequence. One dimension codes the items with respect to the beginning of the sequence using a primacy gradient (the start marker), while the other anchors items to the end of the list using a recency gradient (the end marker); see Figure 11. As the relative strength of the markers is more important than the absolute strength of the markers, the start marker strength at position 1, S_0, was fixed at 1. This reduced the number of parameters required in the model. The value of the start marker for position i is given by:

$$s(i) = S_0 S^{i-1} = S^{i-1} \tag{1}$$

where S is between 0 and 1, and reduces the start marker strength over each position. The strength of the end marker for position i is given by:

$$e(i) = E_0 E^{N-i} \tag{2}$$

where $E_0 > 0$, and usually $E_0 < 1$ in order to produce a smaller recency effect than primary effect, as is seen empirically. E is between 0 and 1, and N is the number of items in the list, in our case 7.

Sequences are recalled by reinstating the start and end marker values for successive list positions, and retrieving an item in response to each cue (i.e., each combination of start and end markers). Each item on the list is activated to an extent determined by the overlap (o):

$$o(p(i), p(j)) = \{p(i) \cdot p(j)\}^{1/2} \exp\left\{-\left(\sum (p_k(i) - p_k(j))^2\right)^{1/2}\right\} \tag{3}$$

between the positional representation with which it was paired, $p(i)$, and the cued positional representation, $p(j)$. In Equation 3, $p(x)$ is a vector representing the positional codes (i.e., a 2-element vector composed of a start marker value and an end marker value), and k indexes the two (start and end) components of each positional representation. Item markers are more distinct at the extreme serial positions, leading to fewer positional errors, resulting in the expected primacy and recency effects and the expected pattern of transposition errors (when items are recalled at the wrong position).

In the original SEM, recalling an item when cued with a particular position marker was accomplished using a noisy winner-takes-all mechanism. Here, the aim was to determine the probability of recall of particular items without requiring Monte Carlo simulation, so instead a modified Luce choice rule (Luce, 1963) was used to convert the activations into recall probabilities. The probability of recalling the item at position i when cued with position j is given by:

$$P(i, j) = \frac{o(p(i), p(j))^{\lambda}}{\sum_{i=1}^{N} o(p(i), p(j))^{\lambda}} \tag{4}$$

where $\lambda \geq 0$ adjusts the sensitivity of recall probabilities to variation in item activation. If $\lambda = 0$, the probability of recalling an item does not depend on its activation at all, and all items will be equally likely to be recalled when cued with any position.

The predictions of the basic SEM, as just described, were obtained as a baseline

indicator of the performance of a model that did not incorporate any long-term constraints.

Implementation of syntactic knowledge

The nature of syntactic constraints means that all items on the list need to be taken into account when recalling individual items. Although this is possible in an item-by-item recall mechanism, it is both conceptually and computationally simpler to instead calculate the probability of an entire sequence being recalled in a particular order (e.g., Dennis, 2009). Accordingly, for each of the 7! potential recall orders in our experiments, we calculated the probability of recalling a sequence in that order as the product of the individual recall probabilities of the constituent items in their given positions. As a consequence, we did not explicitly implement a response suppression mechanism (whereby items which have been recalled have their accessibility squashed at later output positions; Farrell & Lewandowsky, 2002; Henson, 1998; Page & Norris, 1998). This was instead accomplished by constraining reportable sequences to those containing no repetitions, which effectively implements the reconstruction procedure used in the experiments. We did produce an alternative version of the model which included item-by-item recall with competitive queueing and response suppression; the overall pattern of results was very similar to the models described here, and we report the simpler version in this paper. Results of the simulations implementing competitive queueing are presented in the supplemental materials.

A *biasing* version of the model—in which redintegration is the mechanism by which long-term knowledge constrains recall—was implemented by assuming that entire sequences differ in their prior probability of recall according to the frequency of occurrence of the whole sequence in the English language. Frequency of occurrence for the pattern of form classes (e.g., nouns, verbs, adjectives) constituting a 7-word sequence was estimated from the British National Corpus (British National Corpus Consortium, 2007). While whole-sequence frequencies may be the main constraint on ordering, it is also possible that more local pair-wise relations are the basis of the effects

seen in the experiments. A second biasing model is reported using the pairing frequencies analysed in the experiments to calculate prior probabilities. As described in Experiment 1, the frequency of pairs of word classes within each sequence is averaged to produce an overall pairing frequency for the whole sequence. Specifically, for each sequence the prior probability of recall was calculated as:

$$P(s) = kf^{\gamma} \tag{5}$$

where f is whole-sequence frequency or pairing frequency described above, $\gamma > 0$ adjusts the sensitivity of the prior probabilities to variation in frequency of occurrence, and k is a constant chosen such that the sum of $P(s)$ across s is 1. Rather than estimating k, we simply determine $P(s)$ by taking the values given by Equation 5, assuming $k = 1$, and dividing each value by the sum of all values to produce a probability distribution. If γ is set to 0, the prior probabilities of responding with each sequence are equal, and the model acts like the original SEM, with no sequential constraints. The posterior probability of recall of each sequence, based on the prior probability and the probability of recall from the mechanisms in SEM (i.e., the likelihood), was then obtained by multiplying the prior and likelihood values for each sequence, and again dividing by the sum to ensure the posterior probabilities added to 1 (Dennis, 2009).

It is less obvious how to implement *chunking* in the framework of existing models of serial recall. Initially, we implemented a chunking version of the model by assuming that sequences are grouped at input according to their syntactic structure. That is, it was assumed that each parsed token was encoded as a separate group in SEM. SEM, like a number of serial recall models (Brown et al., 2007, 2000; Burgess & Hitch, 1999, 2006; Farrell, 2012; Lewandowsky & Farrell, 2008) accounts for the varied effects of grouping on serial recall (e.g., Farrell & Lewandowsky, 2004; Henson, 1999; Hitch, Burgess, Towse, & Culpin, 1996) by assuming a hierarchical representation of position. Specifically, SEM assumes that one start-end marker pair codes for position within a group, while a second pair codes for the position of groups in the entire sequence; see Henson (1998) for more detail. By encoding chunks as groups, it was thought that the

model could capture the facilitating and regularising effects of syntax. However, we found that this model predicted increasing accuracy with a greater number of smaller chunks for a large range of parameter values (presumably due to greater distinctiveness between items), and therefore did not capture the spirit of syntactic chunking that we had envisaged.

Instead, we modelled a version of chunking that was agnostic as to the specific process by which chunking is achieved. This chunking model was essentially the same as the biasing model described above, but such that the prior probabilities of each potential response sequence were determined according to whether they contained syntactic chunks from the presented sequence. By default, prior values were set to 1. Two versions of chunking are reported: the first involves boosting the prior value of response sequences when a syntactic chunk from the presented sequence appears in exactly the same position in the response; the second involves boosting of prior values when a syntactic chunk from the presented sequence appears in *any* position in the response. Accordingly, sequences are more likely to be produced if they contain more similar phrase structures (noun phrases, verb phrases) to those in the presented sequence.[1] To produce posterior probabilities, prior probabilities were multiplied by the likelihood of producing the particular recall order for each sequence in SEM, and normalised by dividing by the sum of all the posterior probabilities, as described for the biasing model above.

Modelling results

We were interested in the extent to which the chunking and biasing models predicted the dependence of accuracy on the number of parsed tokens at input, and the bias to produce sequences with fewer tokens. For each model, predictions were generated using a reasonable range of parameter values, as presented in Table 9. The search of parameter space represented a full factorial crossing of these parameter values.

[1] We produced models either adding the length of the matched chunk to the prior value for each match; or alternatively adding 1 to the prior value for each matched chunk. These models produced similar results, so only the models boosting by the length of the matched chunk are reported here.

For each set of parameter values, each model was presented with the actual sequences presented to participants in Experiments 1 to 4, and four summary output measures were calculated from the model predictions: a) the average accuracy of recall; b) the slope of the relationship between accuracy and number of parsed tokens in the input; c) the mean change in the number of parsed tokens between the presented sequence and the recalled sequence; and d) the slope of the relationship between the mean change in the number of parsed tokens between the presented sequence and the recalled sequence, and the number of parsed tokens in the input. To calculate these measures, we needed to turn the posterior probabilities for each sequence that were predicted by the models into aggregate measures. Model predictions for metrics (a) and (c) for each sequence presented to each participant were obtained by calculating a weighted average across all possible recallable sequences, the weights being the posterior probabilities of those response sequences. In other words, the accuracy of each potential response sequence was scored, and those accuracy values were averaged across sequences, weighted by the posterior probabilities of the response sequences according to the model. Having obtained predictions for average accuracy and change in parsed tokens on each trial according to the model, we could then construct metrics (b) and (d) using the number of parsed tokens in the presented sequences.

Figures 12, 13, 14, and 15 show the resulting profile of predictions from the models, with the 95% confidence intervals from the empirical data indicated by a cross (these plots were inspired by a similar presentation in Howard, Jing, Rao, Provyn, & Datey, 2009). The top row in each figure shows the predictions of the original SEM model; the second and third rows show the predictions of the biasing models based on whole-sequence frequencies and pairing frequencies respectively; and the fourth and bottom rows show the predictions of the chunking models with boosting for chunks in the correct position and chunks in any position, respectively. Each point represents the predictions under a particular parameter value set. Some sets were excluded because that combination of parameters produced performance at ceiling [p(correct) $>$ .9] or at floor [p(correct) $<$ .1, given chance values of 1/7], and in such situations little effects of

syntax would be expected.

Figures 12 to 15 can be interpreted in several ways. First, the spread of the points gives an indication of the range of different outcomes a model is able to produce based on variations in parameter values. If the dots cover a wider range of space, this indicates that the model can produce more different quantitative predictions, and is an indicator of the flexibility of the model. It is also informative to note where in the space the dots are clustered; for example, all of the models incorporating some syntactic constraints tend to produce accuracy slopes less than 0 (i.e., the points bunch to the left in the middle panels). Finally, an indication of how well a model accounts for the observed data is given by the relationship between the black cross (the data and their 95% confidence intervals) and the model's predictions. A model that predicts the data will contain the black cross inside its cloud of grey predictions, and the extent to which the predictions cluster close to the data indicates how well the model predicts the data overall.

Inspecting the plots from these perspectives highlights a number of patterns in Figures 12 to 15. The first is that the original SEM model without any chunking or biasing mechanism (top row in each figure) cannot predict the empirical results (shown by the cross-hairs). For each experiment, the original SEM model predicts no impact of syntax, so that there is no change in accuracy with parsed tokens condition, and therefore the 'Accuracy Slope' is always zero (the vertical array of dots in the top middle and top right plots of each figure). There is some small variation in the average change in parsed tokens (top left and top middle plots), which is negatively correlated with accuracy (higher accuracy leads to a smaller change in the number of parsed tokens). The variation is presumably driven by a 'regression to the mean' on the number of parsed tokens, which increases when accuracy is lower.

The second pattern is that the biasing model based on whole-sequence frequencies (second row of each figure) can produce a much more negative average change in number of parsed tokens (the ordinate in the left and middle columns) compared to any other model. Third, both chunking models, but particularly the version that only

boosts for syntactic chunks that appear in exactly the same position, struggle to match the combination of accuracy slope and average change in parsed tokens from the data (middle panels). Overall, the data seem to be most consistent with the biasing model based on pairing frequencies, for which the model predictions bunch most closely to the data.

A sense of the accuracy of the models in predicting the data is given by counting how many of the 32,768 parameter combinations for each model produce results within the 95% confidence intervals across all four of our outcome measures of interest. Table 10 shows the number of combinations of parameter values that produce results falling within the confidence limits for all four measures. (Note that we do not treat the confidence intervals as devices for inference, but simply use them as a heuristic for which predictions fall reasonably close to the data, and in a way that is sensitive to the sampling variability in the data). Table 10 shows that the biasing models are more likely to produce predictions that fall within the defined distance of the data than the chunking models or the original SEM model. In fact, the only chunking model that produced any predictions that approximated the data was the model that boosted matching syntactic chunks appearing in any position in the responses, and this was only for Experiment 3.

To be sure that these results demonstrated properties of the models and not the particular parameter values chosen, we re-ran the pairing frequency biasing model and the chunk-anywhere model for Experiment 3 with an extended set of values for the γ parameter. The values used for the other parameters remained as in Table 9. Given that the few fitting values for the chunk-anywhere model had a γ of 1.1, we increased that model's chances of fitting by including more values close to 1.1, as follows: .8, .9, 1, 1.1, 1.2, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5. This resulted in 13 combinations of parameter values that gave predictions falling within the confidence intervals of the data for the chunk-anywhere model, and 739 such combinations for the pairing frequency biasing model. This gives some reassurance that the chunking models fail to capture the data simply because the models were not made to be sensitive enough to the syntax

information.

In summary, all four models incorporating syntactic information were able to account for the qualitative effects seen in the data: a negative change in the number of parsed tokens, a negative accuracy slope, and a negative slope of change in the number of parsed tokens. The biasing model based on pairwise frequencies of grammar classes showed the overall least flexibility, such that its predictions were least sensitive to changes in parameter values. Nonetheless, its predictions tended to bunch most closely to the data, indicating that its predictions were, on average, more in line with the observed data.

In closing this section, we note that in conducting the simulations, two other short-term memory models were explored: the SIMPLE model (Brown et al., 2007) and a stripped down version of the syntagmatic-paradigmatic model of Dennis (Dennis, 2005, 2009). The SIMPLE predictions were excluded from consideration as that model consistently produced the wrong accuracy pattern in predicting better accuracy for a larger number of parsed chunks (worse conformance to syntax), similar to the SEM grouping model. This was essentially due to greater distinctiveness along the grouping dimension with more groups. The biasing version of the Dennis model appeared to behave quite similarly to the SEM version presented here; a decision was taken not to continue using that model as it was not clear how to straightforwardly implement a version of the chunking mechanism.

General Discussion

This paper investigated syntactic constraints in verbal short-term memory, and the evidence for those constraints operating through chunking or redintegration. There were three main patterns in the results. First, greater conformance to syntactic rules led to more accurate recall of sequences in general. Second, given the presented sequences, response sequences were more syntactic and more consistent with the statistics of natural language (as represented by the British National Corpus) than would be expected by chance, representing a systematic bias in recall. Third, response

latencies provided limited evidence for chunking according to hierarchical syntactic phrase structure rules.

Recall accuracy generally increased with conformance to syntax, which is consistent with previous findings (Baddeley et al., 2009; Jefferies et al., 2004; Miller & Isard, 1963; Miller & Selfridge, 1950; Perham et al., 2009) showing that memory for sequences improves as they become more like standard English. Many such studies compared complete sentences to unrelated words (Brener, 1940; Jefferies et al., 2004), but did not investigate the effect of syntax within non-sentences. The current results corroborate the findings of Perham et al. (2009) that syntactic structure within non-sentences enhances recall, and extends them to show that increasing the degree of syntactic regularity within non-sentences has a monotonically increasing impact on memory accuracy.

Whilst this linear trend was not ubiquitous, it was present in three out of four experiments (with Experiment 2 being the exception). However, it should be noted that the effect of conformance to syntax on accuracy was not particularly strong throughout the experiments. For example, in Experiments 1 and 3, where the effect was largest, accuracy was 9% higher in the 2 parsed tokens condition versus the 6 parsed tokens condition, equivalent to an extra 0.6 words correctly recalled on average. Accuracy of recall for word sequences with very little syntactic structure was therefore not much worse than recall of highly syntactic sequences. This was the case even in Experiment 4, where a large number of the sequences had a full sentence structure, highlighting the syntactic nature of the sequences. The relatively modest effect of regularity on recall accuracy is difficult to reconcile with theories that assume that verbal short-term recall is entirely subserved by processes or systems dedicated to perception and language (e.g., D. M. Jones, Hughes, & Macken, 2006; Perham et al., 2009). Although these models have not been instantiated in any detail and thus cannot be ruled out, such models should presumably predict a sizeable impact of conformance to syntactic rules, as more regular sequences will be more compatible with the deployment of language comprehension and production processes. G. Jones and Macken (2015) suggest that the

common co-occurrence of random groupings of digits leads to increased recall in digit span compared to word span, and indicates a role for prior learning of associations between items in short-term recall. For this reason, we might expect little impact on recall for different orderings of digits, but a much larger effect for different orderings of words, which we do not see in our experiments. It should be noted that (Baddeley et al., 2009) found a stronger impact of sentence structure compared to less structured word lists, using a serial recall task with or without concurrent processing. (Baddeley et al., 2009) observed that a major benefit to sentences came from a reduction in order errors, in line with our own paradigm where item errors were prevented. It is possible that some of the benefits of syntactic knowledge obtain for item memory (e.g., Allen, Hitch, & Baddeley, 2018), and that our testing of ordering alone—and requiring items to be recalled in the correct absolute position—reduced the size of our accuracy effect. An alternative possibility is that the mixing together of sentences and non-sentences in Experiment 4 discouraged participants from relying on syntactic knowledge to support working memory. Again, this would seem to be consistent only with a model in which such effects are subject to strategic control, but would seem inconsistent with the findings of (Baddeley et al., 2009) that articulatory suppression did not modulate the sentence superiority effect.

While syntax had an affect both on recall accuracy and on biases in recalled sequences, the results of Experiment 2 suggest that there was little impact of the semantic meaning of sequences on short-term recall in the current experiments. However, there was an effect of semantics on biasing: Syntactic biasing was stronger for the meaningful group than the meaningless group when the presented sequences conformed poorly to syntactic rules. This indicates that the semantic manipulation was strong enough to have an effect on biasing, even if it did not affect accuracy. It may be that semantic meaning simply provides an extra cue to the syntactic nature of the sequences when this is not directly clear from the syntax of the presented sequence. The syntactic biasing was always present even when the sequences had reduced semantic meaning (Experiments 2, 3, and 4), and when there were no meaningful sequences in

the whole experiment (Experiments 3 and 4), suggesting that coherent semantics are not necessary for biasing to occur. The meaningless sequences are also informative because they contain permutations of words that are less likely to have been experienced together outside the lab. This suggests that the sequential effects observed here are partly based on the abstract form classes of the constituent words, rather than being based on lexical or semantic relationships.

How do syntactic constraints influence verbal short-term memory?

The observation of syntactic constraints demonstrates that the effects of linguistic knowledge cannot act purely at the level of individual items (G. Jones & Macken, 2015; Lewandowsky & Farrell, 2000; N. Martin & Saffran, 1997; R. C. Martin, Lesch, & Bartha, 1999; Schweickert, 1993), but must act over multiple items in verbal short-term memory (Allen & Baddeley, 2009; Allen et al., 2018; Botvinick & Plaut, 2006; Jefferies et al., 2004; Perham et al., 2009). Contemporary positional models of serial ordering in short-term memory (Brown et al., 2007, 2000; Farrell, 2012; Henson, 1998)—whereby items are retrieved one by one by cuing with their associated position markers—offer no obvious mechanism to account for such sequential effects. In order to account for the effects shown here, such models must be extended to include mechanisms of support from long-term representations acting over multiple items.

Two principle mechanisms were examined that could account for the impact of syntactic constraints on verbal short-term memory: chunking and redintegration. Syntactic chunking during encoding would enhance memory for items which form syntactically valid chunks (e.g., Allen & Baddeley, 2009; Allen et al., 2018; Baddeley, 2000), or have been experienced together frequently in the past. Alternatively, redintegration would involve reconstructing a degraded memory trace (e.g., Botvinick & Plaut, 2006) by comparison to memory traces that were experienced in the past or according to syntactic rules. The presence of syntactic regularisation in people's responses has been taken as indicative of a multiple-item redintegration mechanism (e.g., Botvinick & Bylsma, 2005). It is perhaps less obvious that a chunking mechanism

could lead to syntactic regularisation, as it does not predict any particular bias in errors with respect to syntax. However, as described in the discussion of Experiment 1, a combination of better memory for more syntactic groupings of items (e.g., Perham et al., 2009), and the opportunity for random errors to improve the conformance to syntax of less syntactic groupings, could act like a 'survival-of-the-fittest' mechanism (e.g., Darwin, 1859/1985), resulting in more syntactic responses than might be expected if syntax had no effect on short-term memory. As such, both chunking *and* redintegration could lead to regularisation of word sequences in short-term memory. What is needed is a way to tell between the two mechanisms.

Both mechanisms were implemented as versions of the Start-End Model of short-term memory (Henson, 1998), to examine the predictions of the two types of model across a range of parameter values, and to determine how well those predictions concur with the empirical data. A search of the parameter space for these models showed that redintegration mechanisms could accommodate the empirical data better than chunking mechanisms, at least as we have modelled them here. While the chunking model produced a scattering of predictions that encompassed the data (Figures 12 to 15), it produced a wider range of predictions, so that the predictions were on average less in accord with the observed data. In contrast, the biasing model assuming that recall was constrained by the pairwise frequency of syntax classes produced a tighter clustering of predictions around the data, and so is better supported by the data (Table 10).

The latency data provided mixed evidence for syntactic chunking. In Experiments 1–3, latencies to the first item in each syntactic chunk were no different to latencies to later items in each syntactic chunk, in either the presented sequences or the responses. However, for Experiment 4, which included more syntactic sequences and more participants, response times were significantly slower to the first item in a syntactic chunk (according to the syntax of the response sequences), particularly at serial positions 3 and 4. An aggregate analysis revealed an effect of syntactic position at the third and fourth serial positions, but only according to the chunking pattern in the

response sequences. As a reminder, the presented sequence and the recalled sequence could usually be independently parsed into phrase structures, and extended latencies at chunk boundaries were only seen for chunks defined by the output sequence, and only for positions 3 and 4. On one hand, the lack of an effect across all serial positions is inconsistent with the usual interpretation of syntactic chunking in which a time cost is invoked when transiting between chunks (Anderson & Matessa, 1997; Daily et al., 2001; Johnson, 1972). On the other hand, the results suggest that the deployment of syntactic knowledge is, in some sense, sensitive to the temporal structure of the sequence. One possibility is that syntactic knowledge interacts with the spontaneous grouping adopted by participants, whereby participants may group a 7-word sequence into a 3–4 or 4–3 pattern. Spontaneous grouping might limit or facilitate chunking by syntax, such that chunks are only encoded as such if they match with the grouping structure assumed by participants. This seems to run contrary to the evidence that the beneficial effects of sentential structures on recall obtain relatively automatically Baddeley et al. (2009). Conversely, the latency effects seen may well reflect the deployment of grouping that occurs in line with the syntactic chunks held in memory, as long as the pattern confers to a general 3–4 or 4–3 grouping structure. It is important to note that our latency effects related to the syntax of responses (i.e. the syntax held in memory at recall) rather than the syntax of the presented sequences. This is not entirely in line with the usual idea of chunking as an encoding process. Overall, we consider the latency results to provide a limited support for a chunking mechanism, but with the recognition that chunking may well interact with other mechanisms in serial ordering models—such as grouping—in a more complex fashion.

Overall, the results are more consistent with syntactic constraints operating via redintegration. However, it should be noted that the results and simulations have not settled this issue. All four models incorporating sequential syntactic constraints produced broadly similar qualitative patterns, and—as just discussed—the latency data are potentially compatible with an interaction between chunking and grouping. Indeed, a major challenge here has been that different mechanisms of sequential long-term

knowledge have been very difficult to discriminate. It may be that syntactic mechanisms are involved across encoding, maintenance, and retrieval (e.g., Allen et al., 2018). Beyond the present results, we believe that the simulations urge caution in drawing strong theoretical conclusions from the effects of syntactic or other long-term knowledge on recall in the absence of validation against the simulated predictions of models.

Accounting for sequential constraints in models of serial recall

Our empirical findings are incompatible with standard positional models of short-term memory. How might contemporary models of serial recall be modified to account for these results? Perhaps selection of the next response is influenced by the form classes of the previous few responses (akin to chaining), or else a sequence is stored hierarchically (e.g., Farrell, 2012) and each sub-group is separately redintegrated. Such mechanisms are difficult to accommodate in positional models of verbal short-term memory where items are retrieved one at a time by cueing with their associated position markers (e.g., Brown et al., 2007, 2000; Henson, 1998; Lewandowsky & Farrell, 2008). However, there are potential mechanisms that allow positional models to interact with long-term knowledge of sequences of multiple items. For example, existing models assume a cumulative matching of sequences of items to known chunks (e.g., Burgess & Hitch, 2006), and others assume redintegration of degraded sequences with respect to acquired knowledge of sequences (e.g., Botvinick & Plaut, 2006). Whilst these mechanisms have so far only been implemented to match sequences of specific items (Botvinick & Plaut, 2006; Burgess & Hitch, 2006), it is possible they could be extended to match against the word types (e.g., nouns, adjectives, etc.) by which syntactic rules are defined. Alternatively, our results might be consistent with item–item chaining models (Dennis, 2009; Lewandowsky & Murdock Jr, 1989). Although chaining models have fallen out of favour (Henson et al., 1996; Lewandowsky & Farrell, 2008), evidence consistent with chaining has been observed in some cases (G. Jones & Macken, 2015; Kahana, Mollison, & Addis, 2010). A question for future investigation is whether a chaining model assuming prior associations between frequently paired classes (e.g.,

adjective–noun) could produce the pattern of data seen here.

References

Allen, R. J., & Baddeley, A. D. (2009). Working memory and sentence recall. In A. Thorn & M. Page (Eds.), *Interactions between short-term and long-term memory in the verbal domain* (p. 63-85). Hove, UK: Psychology Press.

Allen, R. J., Hitch, G. J., & Baddeley, A. D. (2018). Exploring the sentence advantage in working memory: Insights from serial recall and recognition. *Quarterly Journal of Experimental Psychology*. Retrieved from `https://doi.org/10.1177/1747021817746929`

Anderson, J., & Matessa, M. (1997). A production system theory of serial memory. *Psychological Review*, *104*, 728-748.

Baddeley, A. D. (1964). Immediate memory and the 'perception' of letter sequences. *Quarterly Journal of Experimental Psychology*, *16*, 364-367.

Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*, 417-423.

Baddeley, A. D., Hitch, G. J., & Allen, R. J. (2009). Working memory and binding in sentence recall. *Journal of Memory and Language*, *61*, 438-456.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge Univ. Press.

Bock, K. J. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*, 355-387.

Botvinick, M. M., & Bylsma, L. M. (2005). Regularization in short-term memory for serial order. *Learning, Memory*, *31*, 351-358.

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review*, *113*, 201-233.

Bourassa, D., & Besner, D. (1994). Beyond the articulatory loop: A semantic contribution to serial order recall of subspan lists. *Psychonomic Bulletin & Review*, *1*, 122-125.

Brener, R. (1940). An experimental investigation of memory span. *Journal of Experimental Psychology*, *26*, 467-482.

British National Corpus Consortium. (2007). *The british national corpus (version 3).* Oxford University Computing Services. Retrieved from `http://www.natcorp.ox.ac.uk/`

Brown, G. D. A., & Hulme, C. (1995). Modeling item length effects in memory span: No rehearsal needed? *Journal of Memory and Language*, *34*, 594-621.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*, 539-576.

Brown, G. D. A., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review*, *107*, 127-181.

Burgess, N., & Hitch, G. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, *106*, 551-581.

Burgess, N., & Hitch, G. (2006). A revised model of short-term memory and long-term learning of verbal sequences. *Journal of Memory and Language*, *55*, 627-652.

Carbon, C. C., & Albrecht, S. (2012). Bartlett's schema theory: The unreplicated âĂIJportrait d'hommeâĂİ series from 1932. *The Quarterly Journal of Experimental Psychology*, 1-26.

Chen, Z., & Cowan, N. (2009). Core verbal working-memory capacity: The limit in words retained without covert articulation. *The Quarterly Journal of Experimental Psychology*, *62*, 1420-1429.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control.* Cambridge University Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87-114.

Daily, L. Z., Lovett, M. C., & Reder, L. M. (2001). Modeling individual differences in working memory performance: A source activation account. *Cognitive Science*, *25*, 315–353.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling

errors. *Communications of the ACM*, *7*, 171-176.

Darwin, C. (1859/1985). *The origin of species. republished 1985.* Penguin Classics: London.

Dell, G. S., & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, *42*, 287-314.

Dennis, S. (2005). A Memory-Based theory of verbal cognition. *Cognitive Science*, *29*, 145-193.

Dennis, S. (2009). Can a chaining model account for serial recall? *The Proceedings of the Thirty First Conference of the Cognitive Science Society*.

Epstein, W. (1961). The influence of syntactical structure on learning. *The American Journal of Psychology*, *74*, 80-85.

Ericcson, K., Chase, W. G., & Faloon, S. (1980). Acquisition of a memory skill. *Science*, *208*, 1181–1182.

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review*, *119*, 223-271.

Farrell, S., & Lelièvre, A. (2009). End anchoring in short-term order memory. *Journal of Memory and Language*, *60*, 209-227.

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review*, *9*, 59-79.

Farrell, S., & Lewandowsky, S. (2004). Modelling transposition latencies: Constraints for theories of serial order memory. *Journal of Memory and Language*, *51*, 115-135.

Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science*, *19*, 329-335.

Gathercole, S. E., Frankish, C. R., Pickering, S. J., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 84-95.

Gilchrist, A. L., Cowan, N., & Naveh-Benjamin, M. (2008). Working memory capacity for spoken sentences decreases with adult ageing: Recall of fewer but not smaller

chunks in older adults. *Memory*, *16*, 773-787.

Gregg, V., Freedman, C., & Smith, D. (1989). Word frequency, articulatory suppression and memory span. *British Journal of Psychology*, *80*, 363-374.

Griffiths, T. L., & Kalish, M. L. (2005). A bayesian view of language evolution by iterated learning. In *Proceedings of the 27th annual conference of the cognitive science society* (pp. 827–832).

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*(3), 441-480.

Hemmer, P., & Steyvers, M. (2009, January). A bayesian account of reconstructive memory. *Topics in Cognitive Science*, *1*, 189-202. doi: 10.1111/j.1756-8765.2008.01010.x

Henson, R. N. A. (1998). Short-Term memory for serial order: The Start-End model. *Cognitive Psychology*, *36*, 73-137.

Henson, R. N. A. (1999). Positional information in short-term memory: Relative or absolute? *Memory & Cognition*, *27*, 915-927.

Henson, R. N. A., Norris, D. G., Page, M. P. A., & Baddeley, A. D. (1996). Unchained memory: Error patterns rule out chaining models of immediate serial recall. *Quarterly Journal of Experimental Psychology*, *49A*, 80-115.

Hitch, G. J., Burgess, N., Towse, J. N., & Culpin, V. (1996). Temporal grouping effects in immediate recall: A working memory analysis. *The Quarterly Journal of Experimental Psychology: Section A*, *49*, 116-139.

Hoffman, P., Jefferies, E., Ehsan, S., Jones, R. W., & Lambon Ralph, M. A. (2012). How does linguistic knowledge contribute to short-term memory? contrasting effects of impaired semantic knowledge and executive control. *Aphasiology*, *26*, 383-403.

Howard, M. W., Jing, B., Rao, V. A., Provyn, J. P., & Datey, A. V. (2009). Bridging the gap: Transitive associations between items presented in similar temporal contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 391-407.

Hulme, C., Maughan, S., & Brown, G. D. A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language*, *30*, 685-701.

Hulme, C., Roodenrys, S., Brown, G. D. A., & Mercer, R. (1995). The role of long-term memory mechanisms in memory span. *British Journal of Psychology*, *86*, 527-536.

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology-Learning Memory and Cognition*, *23*, 1217-1232.

Hulme, C., Stuart, G., Brown, G. D. A., & Morin, C. (2003). High-and low-frequency words are recalled equally well in alternating lists: Evidence for associative effects in serial recall. *Journal of Memory and Language*, *49*, 500-518.

Hurlstone, M. J., Hitch, G. J., & Baddeley, A. D. (2014). Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological bulletin*, *140*, 339.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment?. *Journal of Experimental Psychology: General*, *129*, 220-241.

Jefferies, E., Frankish, C., & Lambon Ralph, M. (2006). Lexical and semantic binding in verbal short-term memory. *Journal of Memory and Language*, *54*, 81-98.

Jefferies, E., Lambon Ralph, M. A., & Baddeley, A. D. (2004). Automatic and controlled processing in sentence recall: The role of long-term and working memory. *Journal of Memory and Language*, *51*, 623-643.

Johnson, N. F. (1970). The role of chunking and organization in the process of recall. *Psychology of learning and motivation*, *4*, 171–247.

Johnson, N. F. (1972). Organization and the concept of a memory code. *Coding processes in human memory*, 125–159.

Jones, D. M., Hughes, R. W., & Macken, W. J. (2006). Perceptual organization masquerading as phonological storage: Further support for a perceptual-gestural view of short-term memory. *Journal of Memory and Language*, *54*, 265-281.

Jones, G., & Macken, B. (2015). Questioning short-term memory and its measurement: Why digit span measures long-term associative learning. *Cognition*, *144*, 1-13.

Kahana, M. J., Mollison, M. V., & Addis, K. M. (2010). Positional cues in serial learning: The spin-list technique. *Memory & cognition*, *38*, 92-101.

Kantowitz, B. H., Ornstein, P. A., & Schwartz, M. (1972). Encoding and immediate serial recall of consonant strings. *Journal of Experimental Psychology*, *93*, 105-110.

Kintsch, W., & Buschke, H. (1969). Homophones and synonyms in short-term memory. *Journal of Experimental Psychology*, *80*, 403-407.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, p. 707-710).

Lewandowsky, S., & Farrell, S. (2000). A redintegration account of the effects of speech rate, lexicality, and word frequency in immediate serial recall. *Psychological Research*, *63*, 163-173.

Lewandowsky, S., & Farrell, S. (2008). Short-term memory: New data and a model. *Psychology of Learning and Motivation*, *49*, 1-48.

Lewandowsky, S., & Murdock Jr, B. (1989). Memory for serial order. *Psychological Review*, *96*, 25.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, p. 103-189). New York: Wiley.

Madigan, S. (1980). The serial position curve in immediate serial recall. *Bulletin of the Psychonomic Society*, *15*, 335-338.

Marks, L. E., & Miller, G. A. (1964). The role of semantic and syntactic constraints in the memorization of english sentences1. *Journal of Verbal Learning and Verbal Behavior*, *3*, 1-5.

Martin, J. G. (1967). Hesitations in the speaker's production and listener's reproduction of utterances1. *Journal of Verbal Learning and Verbal Behavior*, *6*, 903-909.

Martin, N., & Saffran, E. (1997). Language and auditory-verbal short-term memory

impairments: Evidence for common underlying processes. *Cognitive Neuropsychology*, *14*, 641-682.

Martin, R. C., Lesch, M. F., & Bartha, M. C. (1999). Independence of input and output phonology in word processing and short-term memory. *Journal of Memory and Language*, *41*, 3-29.

Mathy, F., & Feldman, J. (2012). WhatâĂŹs magic about magic numbers? chunking and data compression in short-term memory. *Cognition*, *122*, 346-362.

Maybery, M., Parmentier, F., & Jones, D. (2002). Grouping of list items reflected in the timing of recall: Implications for models of serial verbal memory. *Journal of Memory and Language*, *47*, 360-385.

Mayzner, M. S., & Shoenberg, K. M. (1964). Single letter and digram frequency effects in immediate serial recall. *Journal of Verbal Learning and Verbal Behaviour*, *3*, 397-400.

McLean, R. S., & Gregg, L. W. (1967). Effects of induced chunking on temporal aspects of serial recitation. *Journal of Experimental Psychology*, *74*, 455-459.

Mesoudi, A. (2007). Using the methods of experimental social psychology to study cultural evolution. *Journal of Social, Evolutionary, and Cultural Psychology*, *1*, 35-58.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81-97.

Miller, G. A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behaviour*, *2*, 217-228.

Miller, G. A., & Selfridge, J. A. (1950). Verbal context and the recall of meaningful material. *The American Journal of Psychology*, *63*, 176-185.

Naveh-Benjamin, M., Cowan, N., Kilb, A., & Chen, Z. (2007). Age-related differences in immediate serial recall: Dissociating chunk formation and capacity. *Memory & cognition*, *35*, 724-737.

Page, M. P. A., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review*, *105*, 761-781.

Parmentier, F. B. R., & Maybery, M. T. (2008). Equivalent effects of grouping by time, voice, and location on response timing in verbal serial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1349-1355.

Patterson, K., Graham, N., & Hodges, J. R. (1994). The impact of semantic memory loss on phonological representations. *Journal of Cognitive Neuroscience*, *6*, 57-69.

Perham, N., Marsh, J., & Jones, D. (2009, January). Syntax and serial recall: How language supports short-term memory for order. *The Quarterly Journal of Experimental Psychology*, *62*, 1285-1293. doi: 10.1080/17470210802635599

Pinker, S. (1998). Words and rules. *Lingua*, *106*, 219-242.

Poirier, M., & Saint-Aubin, J. (1996). Immediate serial recall, word frequency, item identity and item position. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *50*, 408-412.

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, *29*, 633-654.

Roodenrys, S., Hulme, C., Alban, J., Ellis, A., & Brown, G. D. A. (1994). Effects of word frequency and age of acquisition on short-term memory span. *Memory & Cognition*, *22*, 695-701.

Saint-Aubin, J., & Poirier, M. (2000). Immediate serial recall of words and nonwords: Tests of the retrieval-based hypothesis. *Psychonomic Bulletin & Review*, *7*, 332-340.

Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory & Cognition*, *21*, 168-175.

Smith, N. (2005). Chomsky's science of language. In J. McGilvray (Ed.), *The cambridge companion to chomsky* (p. 21-41). Cambridge, UK: Cambridge University Press.

Stanners, R. (1969). Grammatical organization in free recall1. *Journal of Verbal Learning and Verbal Behavior*, *8*, 95-100.

Tell, P. M. (1972). The role of certain acoustic and semantic factors at short and long retention intervals. *Journal of Verbal Learning and Verbal Behavior*, *11*, 455-464.

Walker, I., & Hulme, C. (1999). Concrete words are easier to recall than abstract

words: Evidence for a semantic contribution to short-term serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1256-1271.

Watkins, M. J. (1977). The intricacy of memory span. *Memory & Cognition*, *5*, 529-534.

Wilkes, A. L., & Kennedy, R. A. (1969). Relationship between pausing and retrieval latency in sentences of varying grammatical form. *Journal of Experimental Psychology*, *79*, 241-245.

Xu, J., & Griffiths, T. L. (2010). A rational analysis of the effects of memory biases on serial reproduction. *Cognitive Psychology*, *60*, 107–126.

Appendix: Examples of verbal materials

The following tables demonstrate some examples of the verbal materials used in the experiments, with *meaningful* sequences presented in Table 15, and *meaningless* sequences presented in Table 16.

Table 1
Average Pearson correlation co-efficient (Average r) between each syntax measure and accuracy, and between each of the syntax measures in Experiment 1. The results of one-sample t-tests investigating whether correlations were significantly different from zero are also shown. Seq = Sequence; Freq = Frequency

Metric 1	Metric 2	Average r	$t(19)$	p
Parsing	Accuracy	$-.16$	4.78	$< .001$
Whole-Seq Freq	Accuracy	.15	5.66	$< .001$
Pairing Freq	Accuracy	.17	8.03	$< .001$
Parsing	Whole-Seq Freq	-.63	34.04	$< .001$
Parsing	Pairing Freq	-.71	42.53	$< .001$
Whole-Seq Freq	Pairing Freq	.56	46.94	$< .001$

Table 2
t-tests comparing the average change in sequence characteristics between presented sequences and responses for the data versus the bootstrap simulation, using the four syntax metrics and the two frequency metrics. $df = 19$ in all cases.

	Average Change in Response	t	p
Parsing	$-.07$	6.33	$< .001$
Whole-Sequence Frequency	$+5.53$	6.5	$< .001$
Pairing Frequency	$+35630$	7.1	$< .001$

Table 3
Results from the within-subjects ANOVAs from Experiments 2, 3, and 4 investigating how semantic nature of the list and serial position affect accuracy.

Variable	Contrast(s)	F	p	η_p^2
Experiment 2				
Semantic Condition	Linear	.4	.535	.021
Serial Position	Linear	74.68	$< .001$	.797
	Quadratic	57.946	$< .001$	.753
Semantic Condition ×	Linear × Linear	.004	.951	$< .001$
Serial Position	Linear × Quadratic	.185	.672	.01
Experiment 3				
Serial Position	Linear	145.603	$< .001$	.885
	Quadratic	60.888	$< .001$	.762
Experiment 4				
Serial Position	Linear	119.4	$< .001$	.754
	Quadratic	231.16	$< .001$	.856

Table 4
Results from the within-subjects ANOVAs from Experiments 2, 3, and 4 investigating how semantic nature of the list and number of parsed tokens affect accuracy.

Variable	Contrast(s)	F	p	η_p^2
Experiment 2				
Semantic Condition	Linear	.4	.535	.021
Parsed Tokens	Linear	1.44	.245	.071
	Quadratic	.05	.826	.003
Semantic Condition ×	Linear × Linear	.005	.945	$< .001$
Parsed Tokens	Linear × Quadratic	.484	.495	.025
Experiment 3				
Parsed Tokens	Linear	12.1	.003	.389
	Quadratic	3.41	.081	.152
	Cubic	9.91	.005	.343
Experiment 4				
Parsed Tokens	Linear	18.9	$< .001$	.327
	Quadratic	11.2	.002	.223

Table 5
Average Pearson correlation co-efficient (average r) between each syntax measure and accuracy, and between each of the syntax and frequency measures in Experiments 2, 3, and 4. The results of one-sample t-tests investigating whether correlations were significantly different from zero are also shown. Seq = Sequence; Freq = Frequency

Metric 1	Metric 2	Average r	t	p
Experiment 2				
Parsing	Accuracy	−.04	1.27	.221
Whole-Seq Freq	Accuracy	.06	2.26	.036
Pairing Freq	Accuracy	.10	3.81	.001
Parsing	Whole-Seq Freq	-.42	24.22	$< .001$
Parsing	Pairing Freq	-.56	41.02	$< .001$
Whole-Seq Freq	Pairing Freq	.71	79.71	$< .001$
Experiment 3				
Parsing	Accuracy	−.09	3.56	.002
Whole-Seq Freq	Accuracy	.02	.48	.64
Pairing Freq	Accuracy	.07	1.93	.069
Parsing	Whole-Seq Freq	-.41	32.56	$< .001$
Parsing	Pairing Freq	-.57	55.38	$< .001$
Whole-Seq Freq	Pairing Freq	.71	99.89	$< .001$
Experiment 4				
Parsing	Accuracy	−.1	5.34	$< .001$
Whole-Seq Freq	Accuracy	.13	5.95	$< .001$
Pairing Freq	Accuracy	.09	4.98	$< .001$
Parsing	Whole-Seq Freq	-.55	91.11	$< .001$
Parsing	Pairing Freq	-.69	88.86	$< .001$
Whole-Seq Freq	Pairing Freq	.69	201.01	$< .001$

Table 6
Results from the within-subjects ANOVAs from Experiments 2 and 3 investigating how semantic condition and number of parsed tokens affect the difference in change in number of parsed tokens between the real data and the bootstrap.

Variable	Contrast(s)	F	p	η_p^2
Experiment 2				
Semantic Condition	Linear	1.995	.174	.095
Parsed Tokens	Linear	3.327	.084	.149
	Quadratic	2.311	.145	.108
Semantic Condition ×	Linear × Linear	7.806	.012	.291
Parsed Tokens	Linear × Quadratic	2.044	.162	.097
Experiment 3				
Parsed Tokens	Linear	8.765	.008	.316
	Quadratic	1.845	.19	.089
Experiment 4				
Parsed Tokens	Linear	37.35	$< .001$	.489
	Quadratic	1.05	.31	.026

Table 7
Results of three within-subjects ANOVA comparing the average change in sequence characteristics between presented sequences and responses for the data versus the bootstrap simulation for Experiment 2, using the parsing metric and the two frequency metrics, and separated by semantic group. $df = 19$ in all cases.

	F	p	η_p^2
Parsing Metric			
Semantic Group	.77	.391	.039
Data vs. Bootstrap	21.59	$< .001$	.532
Interaction	2.00	.174	.095
Whole-Sequence Frequency			
Semantic Group	1.38	.255	.068
Data vs. Bootstrap	16.94	$< .001$	.471
Interaction	.23	.634	.012
Pairing Frequency			
Semantic Group	1.46	.243	.071
Data vs. Bootstrap	27.87	$< .001$	.595
Interaction	.26	.619	.013

Table 8
t-tests comparing the average change in sequence characteristics between presented sequences and responses for the data versus the bootstrap simulation, using the four syntax metrics and the two frequency metrics. $df = 19$ in all cases.

	Average Change in Response	t	p
Experiment 3			
Parsing	−.17	3.84	.001
Whole-Sequence Frequency	+11.50	2.97	.008
Pairing Frequency	+72707	3.19	.005
Experiment 4			
Parsing	−.33	8.81	$< .001$
Whole-Sequence Frequency	+11.50	2.97	.008
Pairing Frequency	+72707	3.19	.005

Table 9
Values of parameters used in the search of parameter space for the chunking and biasing versions of the SEM model.

Parameter	1	2	3	4	5	6	7	8
S	.01	.03	.06	.09	.15	.35	.65	.99
F_0	.01	.03	.06	.09	.15	.35	.65	.99
F	.01	.03	.06	.09	.15	.35	.65	.99
λ	0	.3	.9	1.5	2.5	4	10	15
γ	0	.15	.3	.5	1.1	2	3.2	5

Table 10
Number of combinations of parameter values for each model that generate results within the 95% confidence intervals of the data for each experiment in terms of: average accuracy, slope of accuracy over parsed tokens condition, average change in parsed tokens, and slope of change in parsed tokens over parsed tokens condition

Model	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Original SEM	0	0	0	0
Whole-frequency Biasing	149	658	212	115
Pair-frequency Biasing	203	849	301	149
Chunk-in-Position	0	0	0	0
Chunk Anywhere	0	0	3	0

Table 11
Examples of meaningful *sequences (in rows)*

speedy	healthy	athlete	encourages	eager	young	children
risky	financial	bankers	receive	huge	big	bonuses
fierce	wild	tigers	eat	elegant	springy	gazelles
deep	blue	oceans	contain	beautiful	glowing	creatures
wise	brown	owl	likes	tiny	juicy	mice
evil	galactic	emperor	destroys	gentle	peaceful	planet
strong	invincible	superhero	wears	bright	red	pants
enormous	blue	whale	sings	beautiful	ethereal	song
mad	biological	scientist	breeds	weird	wacky	monsters
short	hairy	hobbit	smokes	smelly	medicinal	pipeweed

Table 12
Examples of meaningless *sequences (in rows)*

peaceful	shrewd	princess	escapes	raucous	airbourne	autograph
round	soppy	maze	wins	cynical	digital	army
sticky	wobbly	clothing	cost	rich	careless	satellites
unlucky	speculative	rivers	creates	wild	parallel	weddings
wooden	brave	shepherds	crave	financial	huge	papers
beautiful	pragmatic	teardrops	escapes	healthy	elected	citizen
glowing	effective	brains	publish	powerful	juicy	pupils
concise	enormous	weather	leaves	heartfelt	disorderly	boy
silky	snappy	pedestrian	taunts	elegant	tenacious	objects
intelligent	galactic	fool	secures	fierce	red	accidents

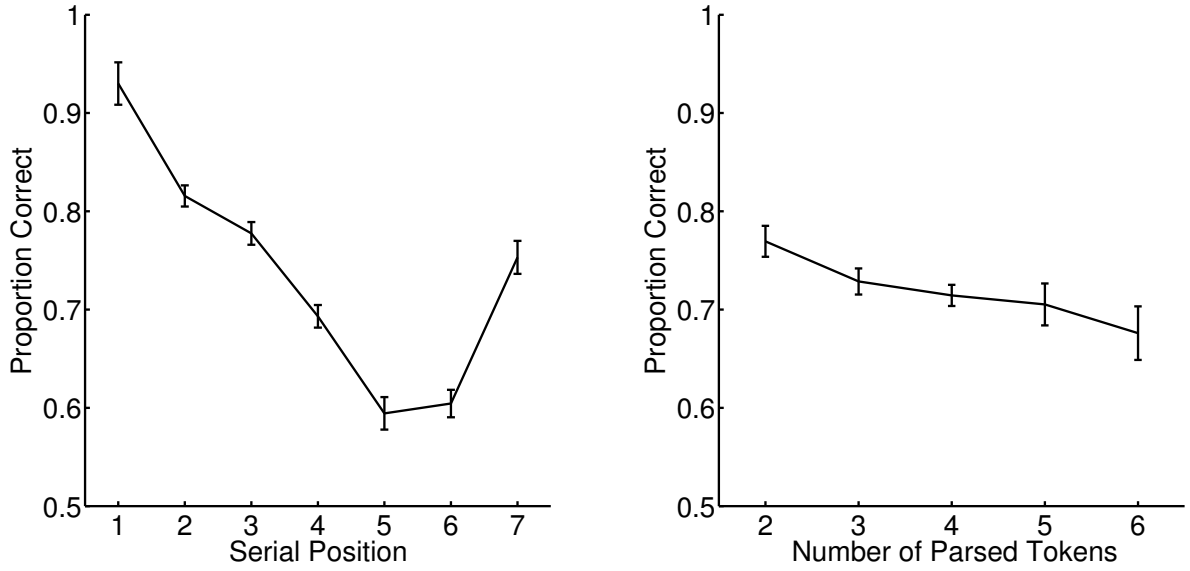

Figure 1. Proportion correct as a function of serial input position (left panel) and number of parsed tokens in the presented sequences (right panel) for Experiment 1. Error bars represent within-subjects standard errors.

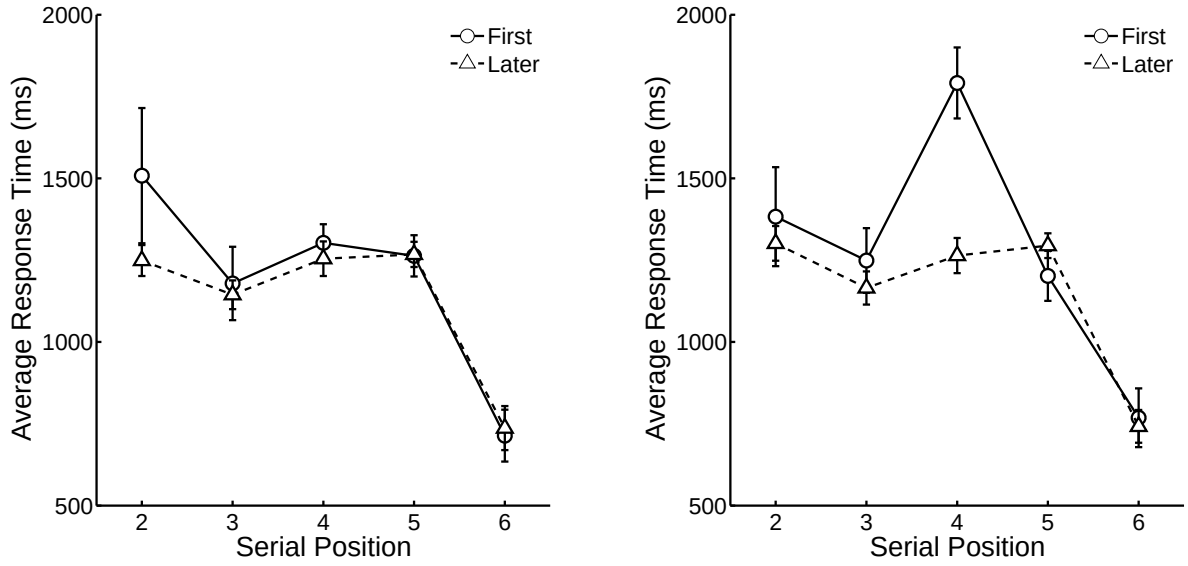

Figure 2. Average response latencies as a function of serial position and the position of an item within a syntactic token (syntactic position) in the presented sequences (left panel) and the response sequences (right panel) from Experiment 1. First = first position; Later = any position except first.

.

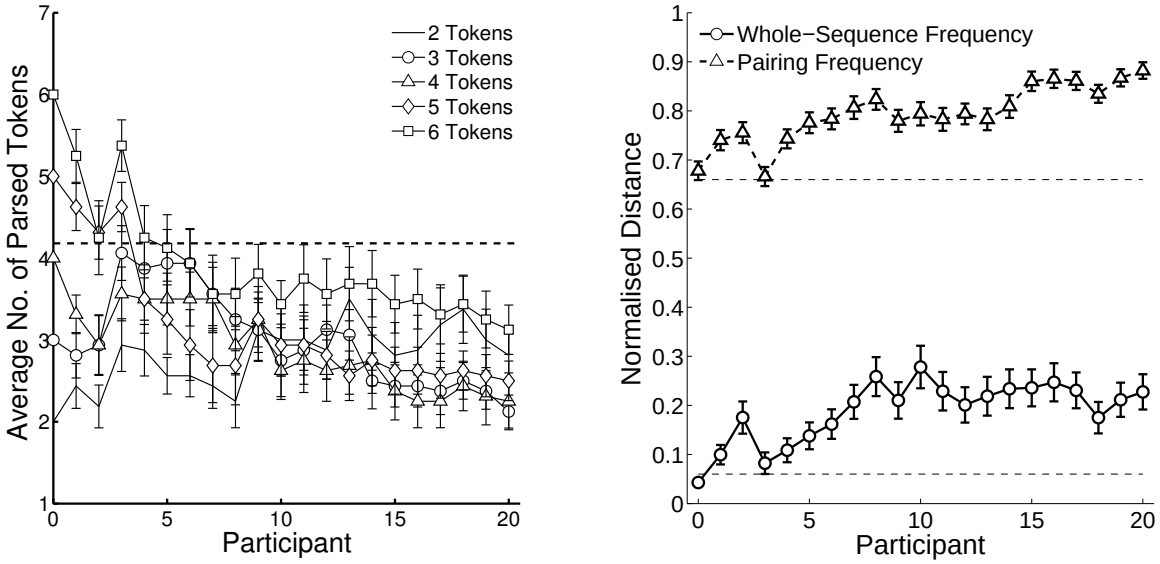

Figure 3. Left: The average number of parsed tokens in sequences that began in different parsed tokens conditions as they are passed along a chain of participants in Experiment 1. Right: The average whole-sequence and pairing frequencies over all of the sequences as they are passed along the chain of participants. The metrics are normalised (divided by the maximum possible for each metric) so that the minimum is 0 and the maximum is 1. Horizontal dashed lines represent the average value of each metric over every possible combination of four adjectives, two nouns and a verb. The data for participant 0 represent the seed sequences generated by the experiment program. Error bars represent standard errors.

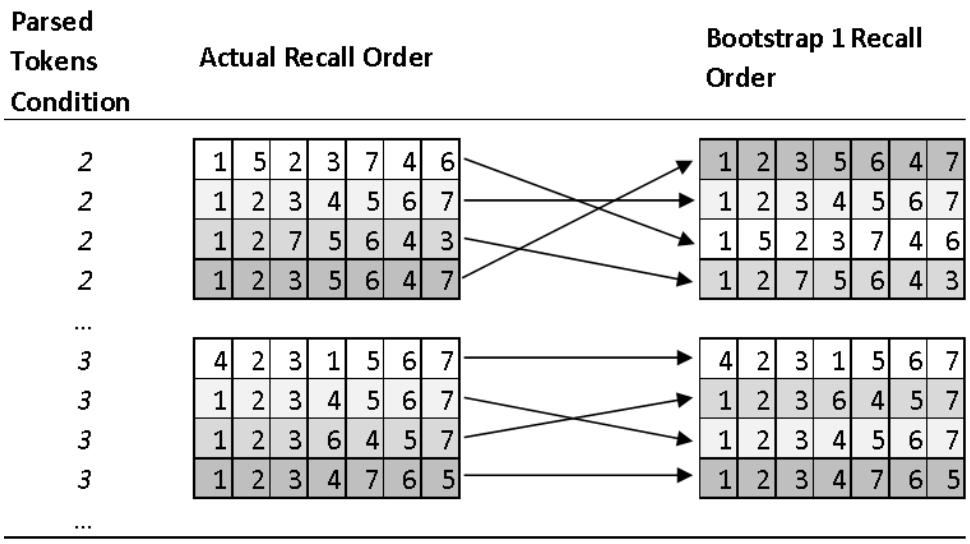

Figure 4. Example of how the bootstrapping process works. This shows only 2 parsed tokens conditions, only 4 trials per condition, and only the first run of 1000 bootstraps, for demonstration purposes.

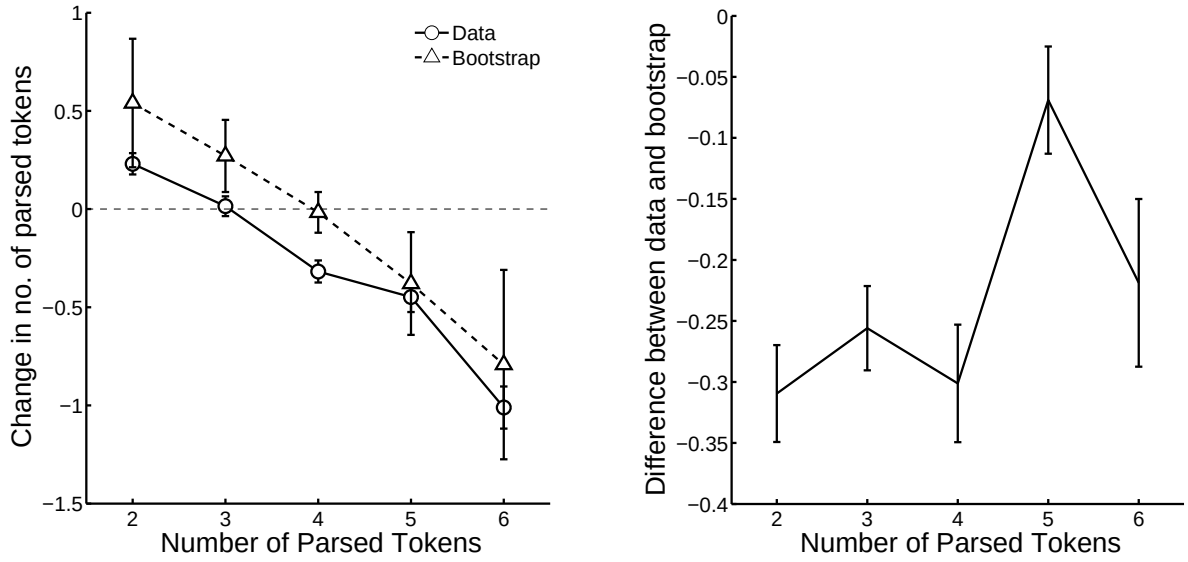

Figure 5. Average change in number of parsed tokens between presented sequences and responses for each participant in the experiment and the bootstrap simulation (left panel) for Experiment 1. Points above the dashed line represent an increase in parsed tokens (decrease in conformance to syntax). The same data are re-plotted to show the difference between the data and the bootstrap for each condition (right panel).

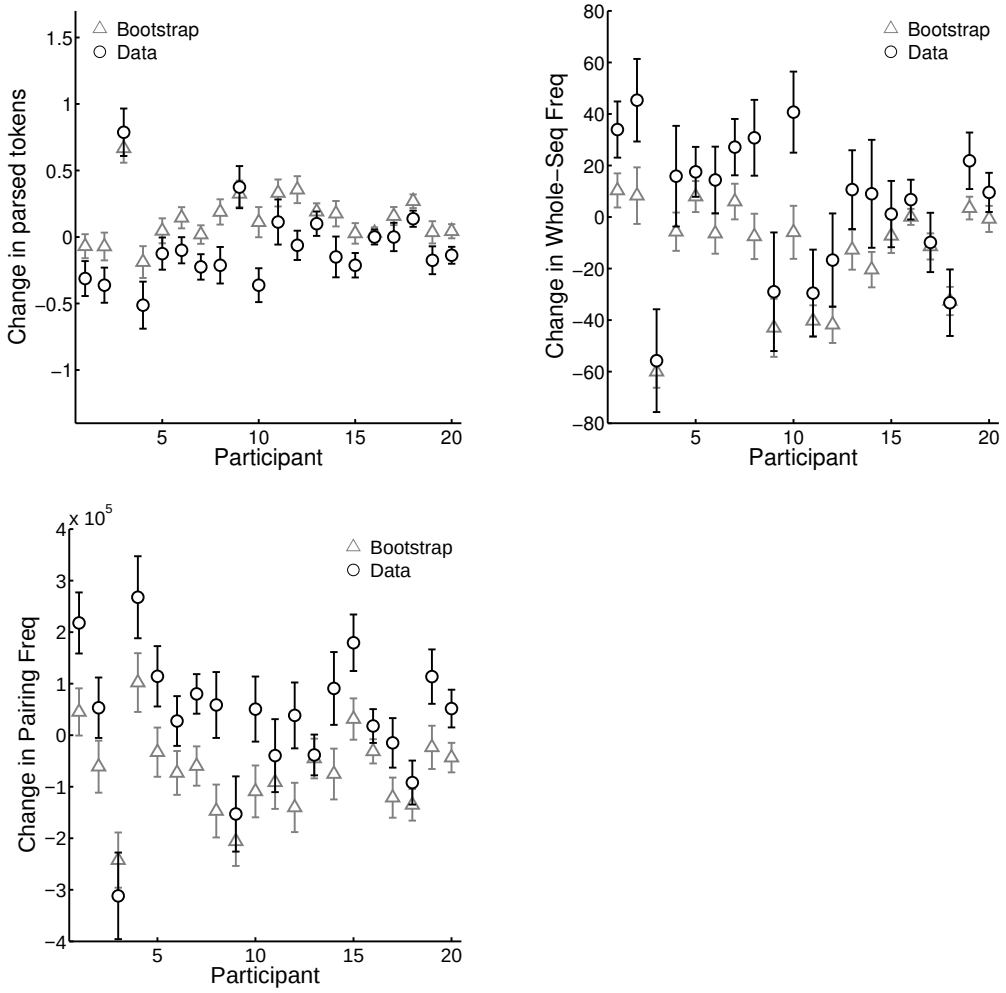

Figure 6. The average change in different sequence characteristics between presented sequences and responses, both in the data and the bootstrap simulation. Top-left: Parsing measure; top-right: Whole-sequence frequency; bottom-left: Pairing frequency. Error bars represent between-trials standard errors.

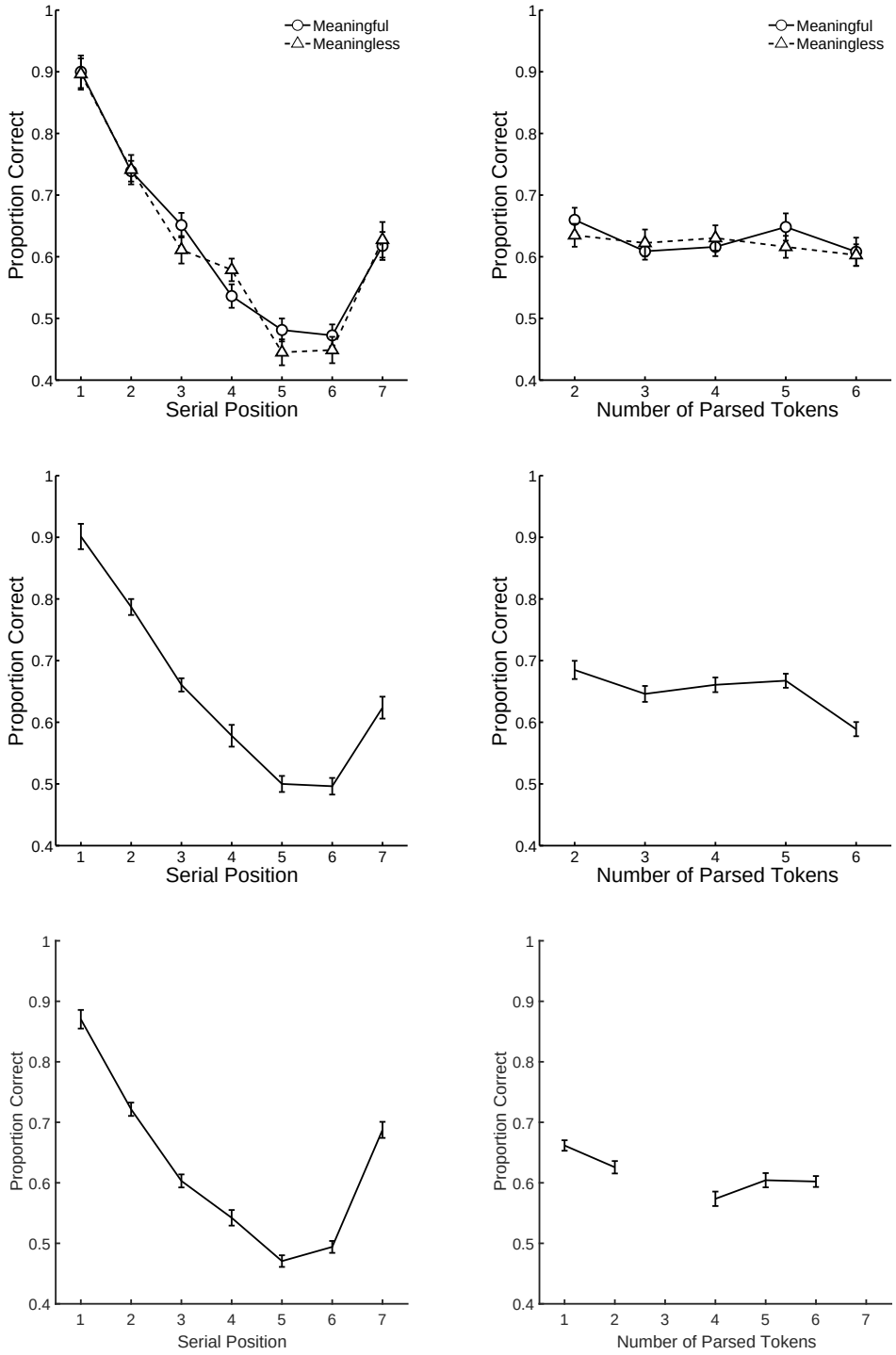

Figure 7. Proportion correct as a function of serial input position in Experiment 2 (top) to 4 (left panels); Proportion correct as a function of number of parsed tokens in the presented sequences for Experiment 2 (top right) to 4 (right panels).

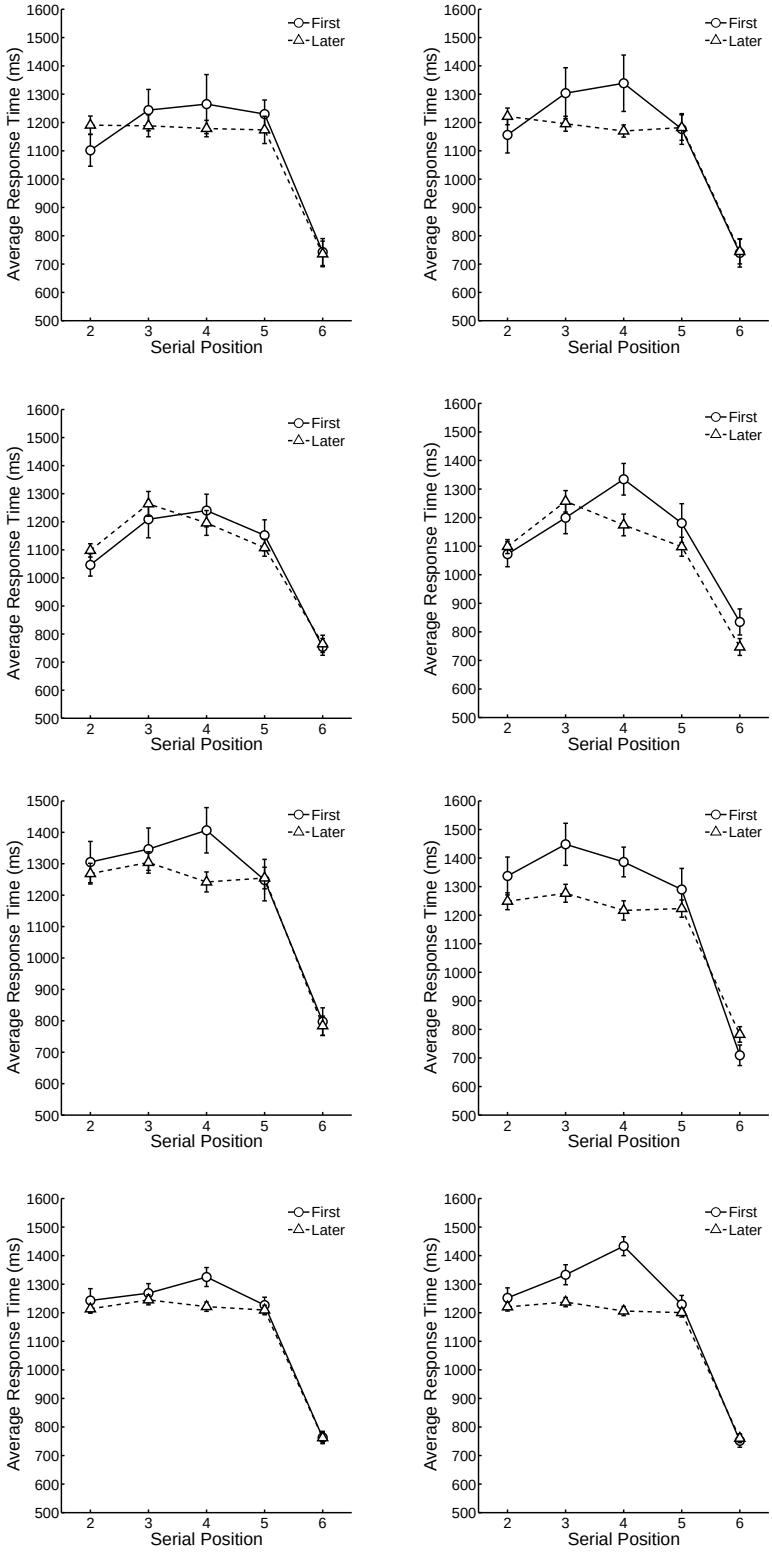

Figure 8. Average response latencies as a function of serial position and the position of an item within a syntactic token (syntactic position) in the presented sequences (left panel) and response sequences (right panel) from Experiment 2 (top), Experiment 3 (second), Experiment 4 (third) the aggregated data over all four experiments (bottom).

.

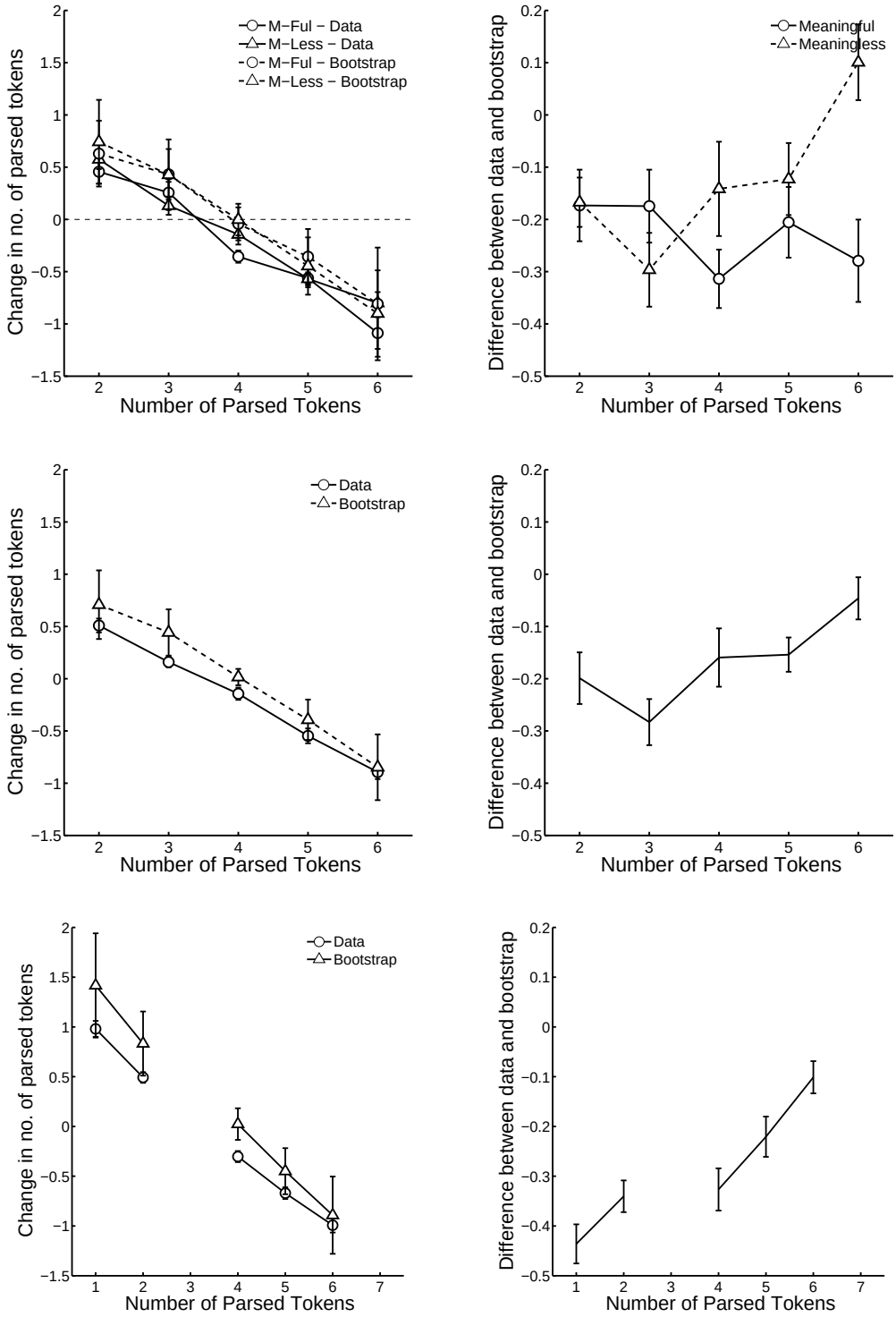

Figure 9. Average change in number of parsed tokens between presented sequences and responses for Experiment 2 (top left), Experiment 3 (middle left) and Experiment 4 (bottom left). The same data re-plotted to show the difference between the data and the bootstrap for each condition (right panels).

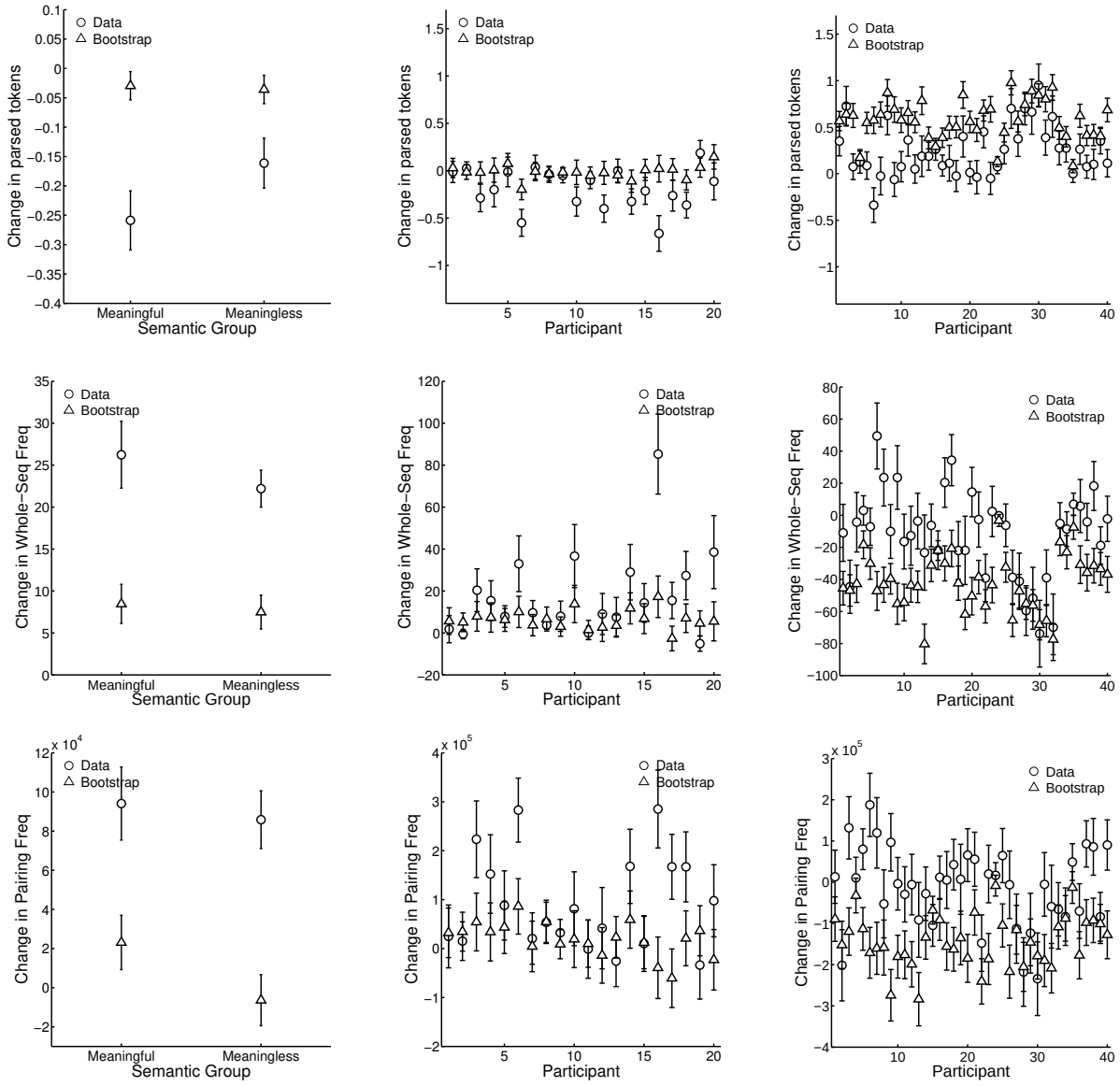

Figure 10. The average change in different sequence characteristics between presented sequences and responses in the data and the bootstrap simulation, for Experiment 2 (left, separated by semantic group), Experiment 3 (middle), and Experiment 4 (right). Top: Parsing measure; Middle: Whole-sequence frequency; Bottom: Pairing frequency. Error bars represent within-subjects standard errors.

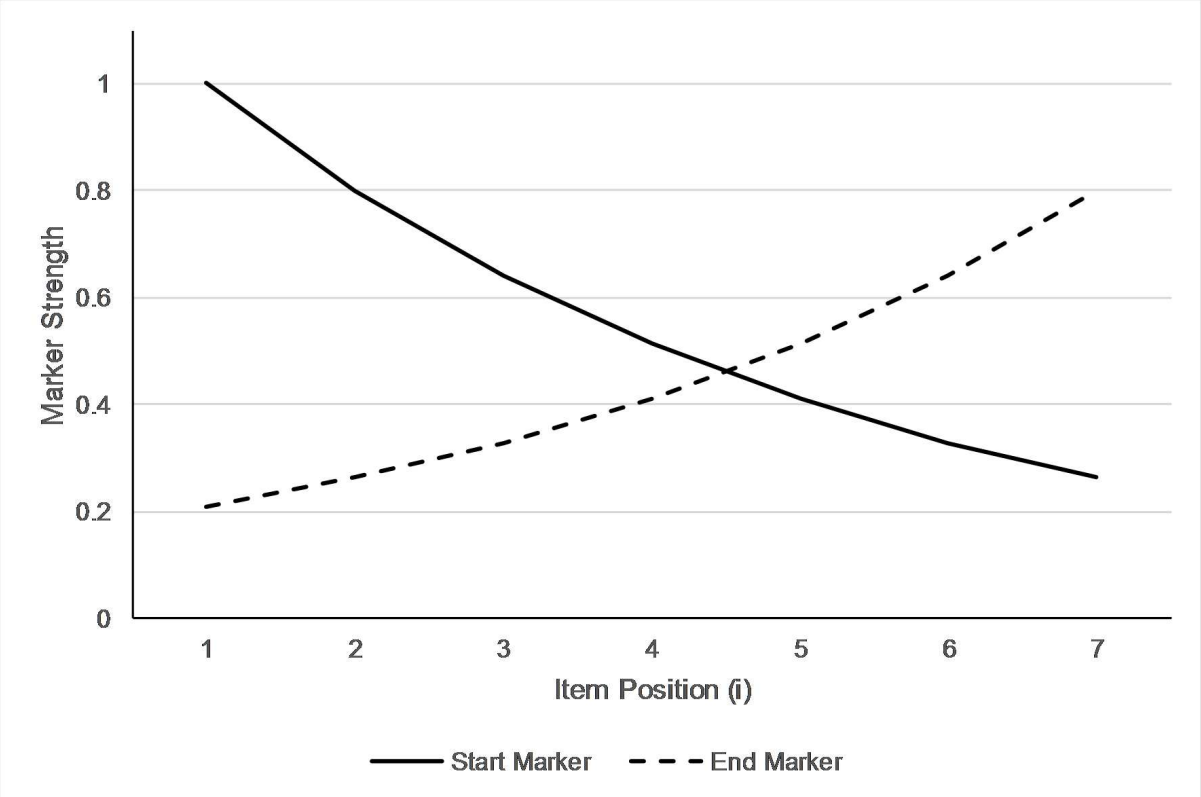

Figure 11. Example showing start and end marker strength at different item positions

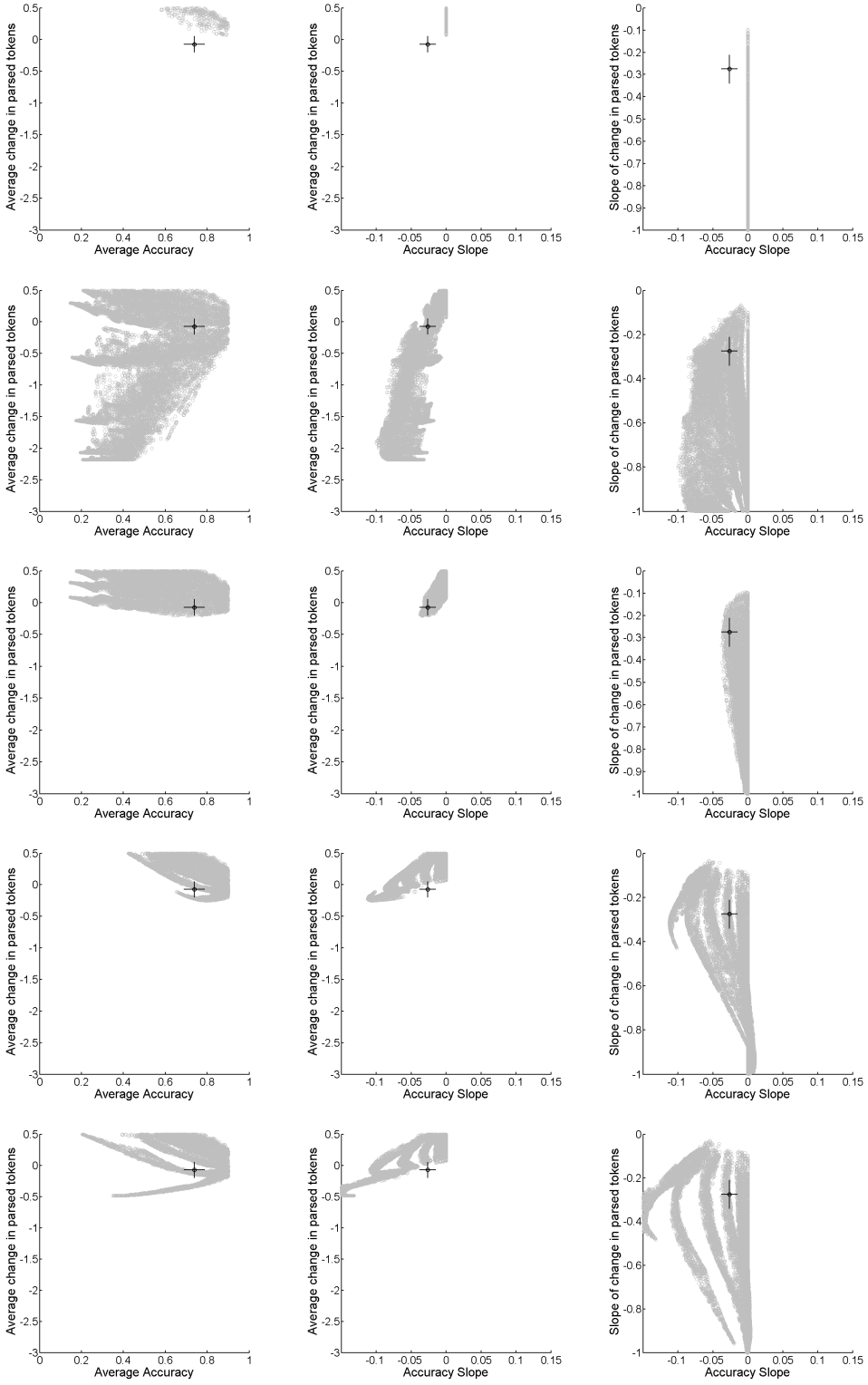

Figure 12. Predictions of various versions of the Start-End model using a range of parameter values, given the presented sequences from Experiment 1. Rows correspond to the original start-end model (top row); whole-sequence biasing model (second row); pairing-frequency biasing model (third row); chunk-in-place model (fourth row); chunk anywhere model (bottom row). The centre of the black cross indicates the average values in the data, whilst the arms of the cross represent 95% confidence intervals on the data.

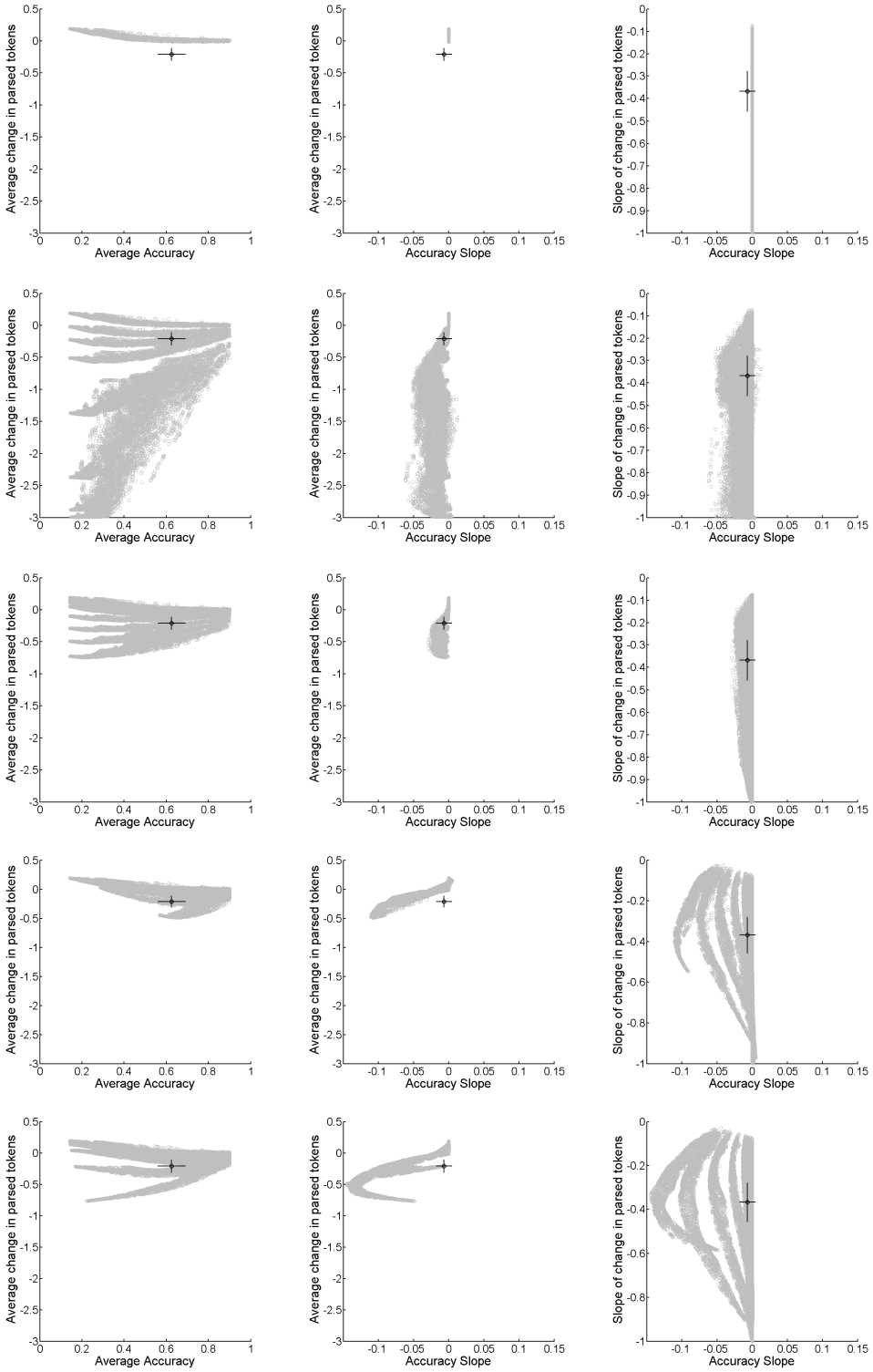

Figure 13. Predictions of various versions of the start-end model using a range of parameter values, given the presented sequences from Experiment 2: the original start-end model (top row); whole-sequence biasing model (second row); pairing-frequency biasing model (third row); chunk-in-place model (fourth row); chunk anywhere model (bottom row). The centre of the black cross indicates the average values in the data, whilst the arms of the cross represent 95% confidence intervals on the data.

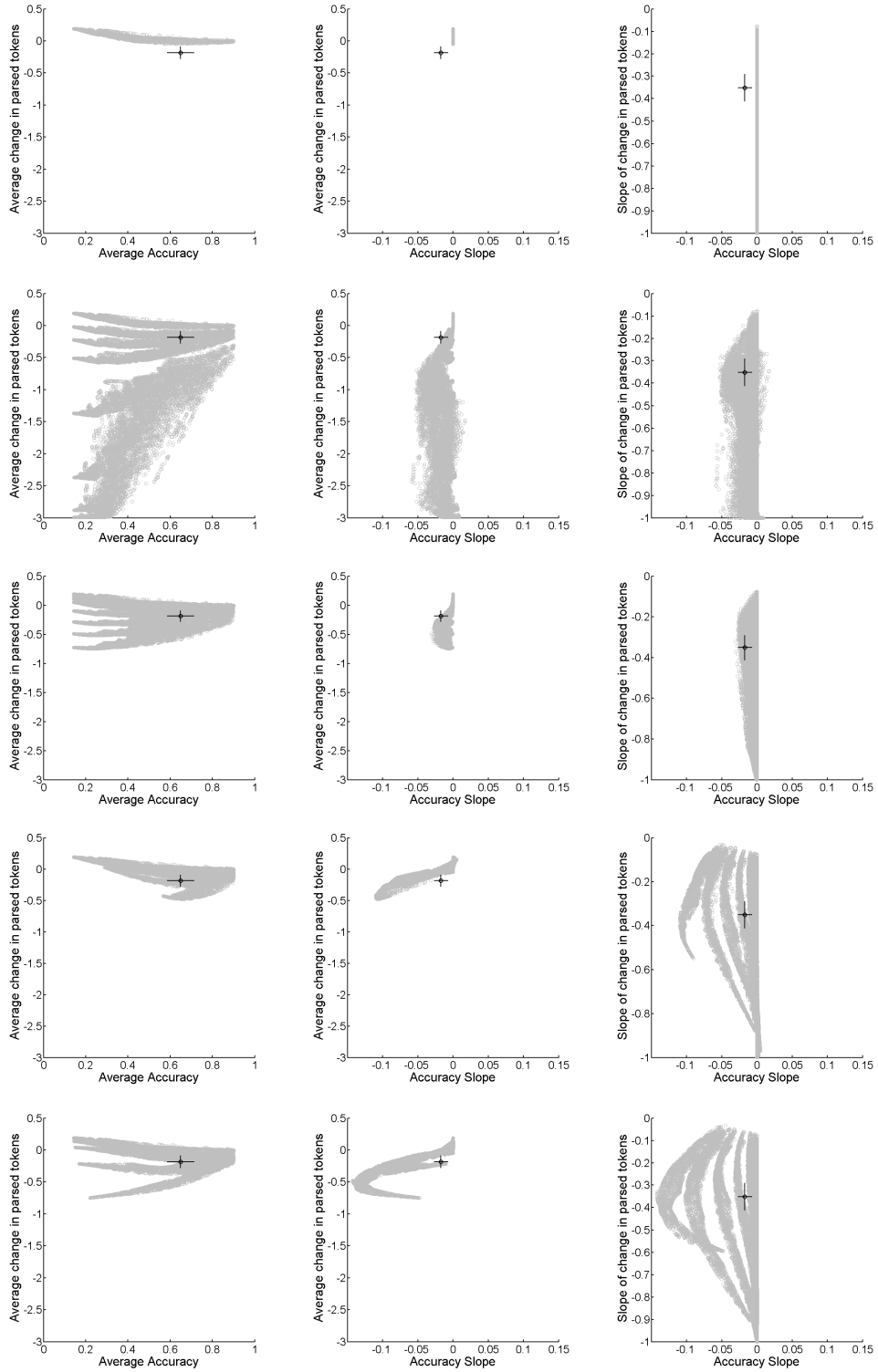

Figure 14. Predictions of various versions of the start-end model using a range of parameter values, given the presented sequences from Experiment 3: the original start-end model (top row); whole-sequence biasing model (second row); pairing-frequency biasing model (third row); chunk-in-place model (fourth row); chunk anywhere model (bottom row). The centre of the black cross indicates the average values in the data, whilst the arms of the cross represent 95% confidence intervals on the data.

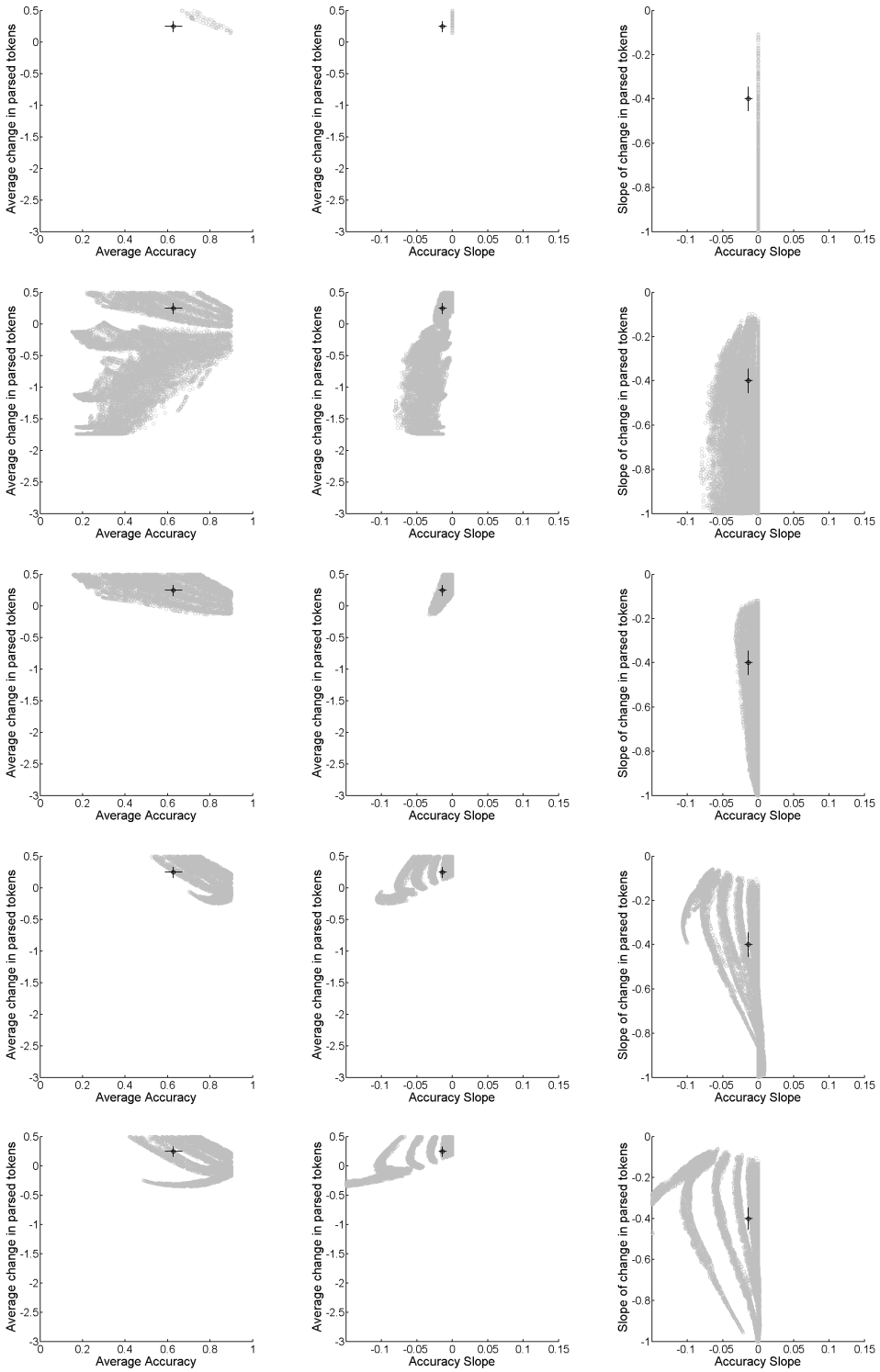

Figure 15. Predictions of various versions of the start-end model using a range of parameter values, given the presented sequences from Experiment 4: the original start-end model (top row); whole-sequence biasing model (second row); pairing-frequency biasing model (third row); chunk-in-place model (fourth row); chunk anywhere model (bottom row). The centre of the black cross indicates the average values in the data, whilst the arms of the cross represent 95% confidence intervals on the data.